WXES 3182
Session 2002/2003

Name : Kheong Ling Ling

Matrix No: WET000065

Supervisor: Mr. Teh Ying Wah

Moderator: Ass. Prof. Dr. Syed Malek Fakar Duani

Discovering Interesting Knowledge in Databases
--Data Mining with Classification

Abstract

This project is prepared to fulfill the requirement of the Bachelor of Information Technology. The development of Implementation of Data Mining is highlighted in this document.

Data Mining as well as knowledge discovery is a solution to overcome data explosion problem. It is an automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.

The main purpose of this research is to find out the useful and power of data mining in the data management. In order to discover the usefulness of data mining, a system of sales-marketing analysis tool had been set up. This can be used to analyze the data from the database and only the related data will be shown as the result. It means that we no need to face all the data by selecting the appropriate one manually. As the result, it can reduce resources and energy on works.

The literature review in the Chapter 2 is a part that gathered all the related system that similar or using Data Mining as well. In this chapter, it consists of the analysis of other systems and yet I can develop a system that is more efficient. Besides, the information about other systems would be a reference to this research as well.

The Waterfall Model with prototyping has been chosen as the software life-cycle model to develop the system. The development technologies and tools are Java, Microsoft Access in the platform of Microsoft Windows 2000.

The report will fully focused on the introduction to the system, literature review, system analysis, and system design till evaluation of system.

Acknowledgement

With the completion of this document, I would like to state my gratitude to several persons who had given a lot of helps. My hearty thanks go to Mr. Teh Ying Wah who is my supervisor for this project. Mr. Teh had given a lots of help and advices for me to come out with this project. Mr. Teh willing to spend his precious time to give a consultant session for us.

Besides, I would like to thank Ass. Prof. Dr. Syed Malek Fakar Duani for being my moderator. During the viva session, Dr. Syed Malek had given some suggestions to my project. Besides, Dr. Syed Malek had pinpointed some parts that needed to make some modification.

My gratitude also goes to Mr. Danny Ong who is a system developer. Mr. Danny willing to make an informal interview with me regarding the system development.

I am also grateful to my fellow group partners. We had spent times to discuss about the topic and the system that need to be developed. After several meetings and negotiations, we had decided to come out with the idea of sales-marketing analysis tool.

Besides, I would not forget to thank my friend who had given their times on testing my system and given opinion as well. There are Phang Siew Ting, Oo Chin Boon and Tay Hsiung Kae.

Lastly, my heartfelt thanks go to my parents and friends who always give me their support and advice. I really appreciate the helps that had been given all this while.

Kheong Ling Ling
February 10, 2003

List of Figures

List of Tables

Table of Content

CHAPTER 3 METHODOLOGY

CHAPTER 4 SYSTEM ANALYSIS

CHAPTER 5 SYSTEM DESIGN

CHAPTER 1

INTRODUCTION

Chapter 1: Introduction

1.1 Introduction

This document outline proposed the plan for the development of the Data Mining system. This is a research on data mining for an over view of the new technology that emerge top nowadays.

The purpose of this introductory chapter is to describe about the general information and knowledge for this project. With this introductory, briefly describe the scope, aims and objectives of the project, project outcome, advantages gain from the system and project schedule. This is to make sure the project can be done on track. Besides, enclosed with this chapter is the layout of the entire project. Data mining is part of Knowledge Discovery Database that we can automate the process of analysis data from the existing databases and make the following decision.

We cannot deny there is an increasing need to have better decision making tools for the huge databases as well as data warehouses. Data mining tools is not as the same as other analysis tool. It had extra advantage when it comes to discover some hidden pattern of data. Thus, discovering the information or beneficial knowledge from the databases has drawn attention from the computer-related field geniuses as well as business traders to help on revealing the hidden or unseen pattern in the data. It does not make sense for collecting the entire data without proper process to extract the useful data for decision-making.

1.2 Project Definition

The title is "Finding interesting knowledge in the databases". Two major points are discussed here which are interesting knowledge and databases. Interesting knowledge here refers to the data that had been processed to become information which further develop to become useful knowledge about historical patterns and future trends. Database is a data store where all the related data are captured. In order to find interesting knowledge in the databases, data mining become the main coverage of this project.

Nowadays, the data overflow and caused the analysts and statisticians facing a lot of difficulties when analyzing data. Traditionally, they have to keep track of the history of the data in order to make the following analysis. This is a time consuming and effort required task. Thereby, with the increase growth of the database, some techniques are needed to improve the data extraction process. This comes out with an idea namely data mining.

By applying data mining techniques, which are elements of statistics, artificial intelligence and machine learning, they are able to identify trends within the data that they did not know existed. Data mining can best be described as a business intelligence technology that has various techniques to extract comprehensible, hidden and useful information from a huge range of data. This technology makes it able to discover hidden patterns and trends in large amounts of data. The outcome of a data mining practice can take the form of patterns, trends or rules that are implied in the data. Through data mining and the new knowledge it provides, individuals are able to categorize the data to produce new opportunities or value for their organizations.

1.3 Project Objective

The increasing of data today leads to the increasing needs to discover the hidden pattern in the data warehouses. We are facing a lot of data everyday, despite of whom, when or where we are located. For instance, in a marketing environment, the data of goods sold and the inventory becomes the main factor of the transaction. If we are to predict how many goods will be sold out in one time frame, the process of filtering relevant data from the huge bunch of data is time consuming. Therefore, this project is done for few objectives as stated below:

- Data mining provides an effective way to extract the data from the database. It's not done manually but it is an automation tool that leads to efficiency of filter off the unrelated data. It generates a model that requires less manually effort.
- The time consuming method of filtering data manually no longer affect the process of getting the data. As it is an automation tool, time being used to evaluate larger number of data is reduced.
- Less technical expertise is needed for the user to utilize the data mining tools. Most filtering and procedures are automated.
- Provides a prediction model and reveal hidden pattern that can be used to find out the potential trend in the future. This is useful for making decision for investment, transaction such as market basket analysis, fraud detection, direct marketing, trend analysis, market segmentation, and interactive marketing.
- To find out the efficiency and accuracy of using data mining technology

- Use Computer Cycles to replace Human Cycles. Instead of doing all the analysis by the personnel, using the computer technology will bring be more effective an efficient.

1.4 Scope of Project

The proposed system is only suitable for the housing developers and intranet's use. There will be 3 modules in general namely, log in module, data testing module, summary module.

For the log in module, user can log in to the system and search for the result of the item and the field they want to analyze. User can either select what category they want to analyze in the analysis column. With this, users can analysis the selected field easily without gone through all the data manually. This is to make sure only authorized user can access to the system.

Another module is data testing. This is the stage when the classified data is being tested in order to find out whether the classification rules are practical for all situations.

Summary module generates the final report of the analysis. It includes charts, reports and analysis.

1.5 Project Schedule

In order to achieve the objectives of the proposed system, a milestone of the whole system is figured out. The milestone arranges the time for each stage and at the same time provides a guideline in developing the proposed system. All the

activities for the project development is written and scheduled in a chart so that sufficient time can be allocated to the activities. In the schedule, all the activities are specified together with the duration that would be spent to implement the project.

Project management is the coordination of all aspects of a project so that it can be completed under the constraints defined.

- Define the goals of the project
- Define and allocate the resources
- Establish timetable and schedules workloads
- Trace and monitor progress of the project
- Report and document the project

Gantt Chart is a formal way to show display the schedule of projects. It is a two-dimension chart with the time frame across the activities. The bar in the Gantt chart shown the period of time activities would be taken. Figure 1 below is a Chat Gantt for my whole project schedule:

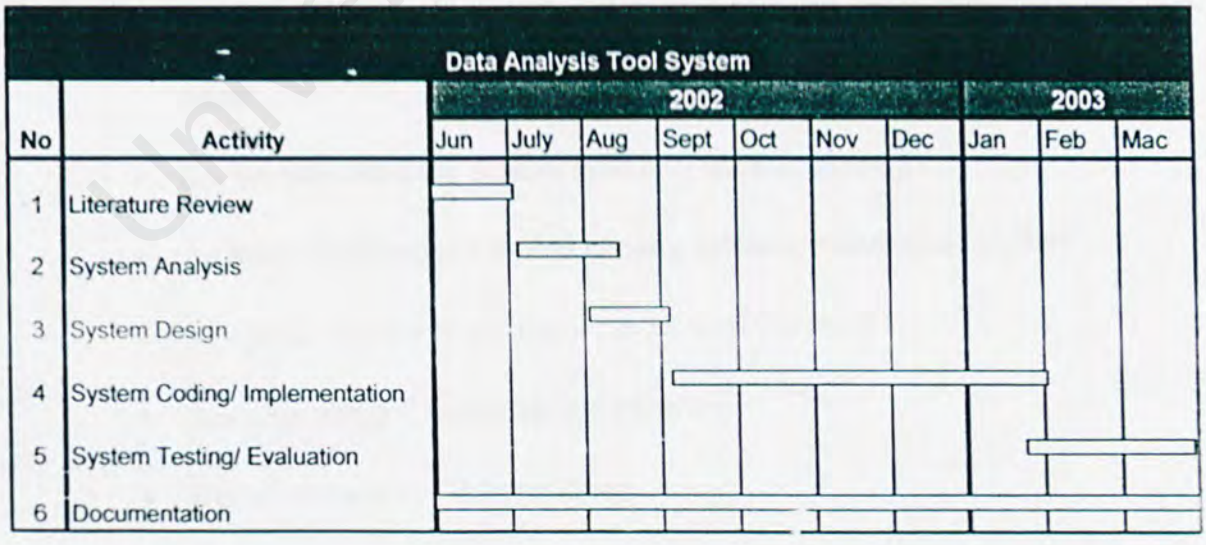

Figure 1.1 Project schedule

From the literature review till the system design, the time allocated is about 3 months. This development will be examined after the system design completed. Then the next is coding or implementation and the allocation due to the complicated that will be brought to. The allocation is 5 months. Last period of one and a half months, the testing of the system will be taken.

The project schedule is prepared to fit it all the stages in the system development. It is a guideline to ensure the system can be delivered on time.

1.6 Expected Outcome

By the end of the project, some outcomes are expected to achieve the objectives of the projects. Below are some expected outcomes:

- An analysis tool of data mining that can generate the knowledge through classification algorithm.
- Analysis with higher accuracy.
- Provides user-friendly interface that leads to better understanding of the usage.
- Let the user input the item or field they want to classify
- An automated analysis tool to classify the data with certain attribute.
- Visual-aid display to get the better view of the result.
- Analysis selected data with less expertise
- Higher accuracy of the prediction.
- Provides user-friendly interface.

1.7 Limitation

Since this is a analysis tool does not require any input from the user, this make the system only support the selected field to be analyzed. Thus, the user just can view the result of the particular technique.

Below are some limitations found in the system:

- Not real-time access
- Only one person can access at one time
- Only one database, thus cannot have large number of data

1.8 Advantages Gained From The System By The Developer

The proposed data mining system is expected to have the following advantages:

a. Reduced time to analysis the data from the huge databases.

b. Capability to provide accurate analysis result with visualization.

c. A hassle free system for users.

d. Automation generates data analysis without much effort.

e. Accurate and up to date information on the database record

f. Only authorized user can access the results to ensure the security of the analysis and confidential of the result.

g. Result-oriented in which the result is the focus.

1.9 Chapter Summary

The purpose of this documentation is to gather all the relevant information that needed to develop the system from the literature review which finding related reference to the analysis of the system, design the system requirements till the end of the maintenance methods. As this document consists of all the requirements and methodology, the consequential steps are based on the steps documented in this paper. As a over view of the report, below is the report layout:

Chapter 1: Introduction

This introductory gives an overview of the projects. From defining the project problem, it followed by the objectives of the projects, project scope, project schedule, expected outcome of the project as well as the advantages gained for the system by the developer.

Chapter 2: Literature Review

This part consists of information that gathered towards researches and studies made on the related field of similar existing system, available programming languages, databases and the appropriate development tools as well. It focuses on the data mining techniques based software as well as analysis tools to explore the strengths and the weaknesses of the system. From the studies, the right tools, techniques, will be chosen to develop the system in order to design the optimize performance that appropriate to them.

Chapter 3: Methodology

This chapter emphasized on the methodology to develop the system. The methods of developing the system have been described in this chapter.

Chapter 4: System Analysis

This chapter pinpoints the requirements of the system that is going to develop. This includes the functional and non-functional requirements of the system. It is the guidance of the development in which the requirements should be fulfilled to produce the output.

Chapter 5: System Design

Chapter represents the conceptual and technical design process of the system such as the considerations for database implementation, functional design and the process design utilized in the system development.

Chapter 6: System Implementation and Testing

This chapter describes the environments of the system's platform after setting up the system. It includes validations and verifications to minimize the level of error.

Chapter 7: System Evaluation and Conclusion

After implementing and testing the system, evaluation will be made to the system in terms of strength and limitations. Suggestions for further enhancements and also the problems encountered during the implementation will be listed here. It ended with the conclusion of the entire project.

CHAPTER 2

LITERATURE REVIEW

Chapter 2: Literature Review

2.1 Introduction

The literature review is an essential part of any academic research project. It contains studies on the existing solution towards the similar problem. Therefore, from the studies of the existing output, we are urged to carry out a new solution other than the existing one. This can be done by find out the strength and weakness of the existing solution following with the new idea to resolve the problems.

The literature review consists of comparison of the data mining analysis tools with the non-data-mining analysis tools, review of the similar existing system, technologies review, operating system, and software development model. With this studies, it will become guidance or so call reference to choose the appropriate languages, development tools, and so does the development model. It helps to the next stages of development process such as system analysis as well as system design.

2.2 Data Mining

2.2.1 What is Data Mining?

Data mining is the extraction of hidden information from databases. It is a powerful new technology to let the companies focus on the appropriate data instead of revising all the data from the data warehouses. Data mining tools can be used to predict future trends and behaviors, thus knowledge-driven decisions can be made by businesses towards their aim to make the right decision.

Data mining tools can resolve business problems that traditionally were too time consuming to resolve. They search databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Generally, data mining is the practice of analyzing data from different perception and summarizing it into practical information in order to make profits and reduce cost. With this, data mining software is created as a tool to analyze data. Data can be analyzed from many different angles or dimensions, in order to classify it, to get the summary of the relationships identified. Data mining can be defined as a process of discovering correlations or trends among massive data in the data warehouse or relational databases.

Analytical tools for data mining are predictive, descriptive or even combination of both. For predictive model, some outcome can be predicted to make knowledge-driven decision. It is based on the existing data that had been classified to the predefined group. Descriptive model gives the correlation of the data. For instance, for selling goods A, probability of the customer to purchase the goods B is high and both products are correlated which means that the products is highly related in the view of marketing.

There is several functionality of data mining that leads to different ways of mining the data. Those are Classification and regression, Association, Clustering Analysis, Sequential Analysis, and

2.2.2 The Scope of Data Mining

In the process of revealing knowledge in the databases, generally, data mining tools provide these capabilities such as: prediction of behaviors and trends automatically and discovery unknown patterns. Data mining allows automation of searching

relevant information from the databases. Since, it is automated, it is obviously had a quick result. Example for the trend prediction is targeted marketing. In this case, data mining allows usage of past promotional mailings data to identify the targeted potential customers. Besides, it can be applied on forecasting the bankruptcy, and forecasting a group of people that would have similar responses or behavior. On the other hand, data mining scan through the databases to reveal unknown patterns beneath. For instance, in the sales of products, that is patterns that products seemingly are purchase together. In the transactions deal, it can be used to identify irregular data to avoid keying errors of entry data.

Data mining techniques can be implemented on high performance parallel processing systems. It can analyze immense data in a short time. It provides fast processing in which several models can be used to reveal the knowledge in the databases.

2.2.3 Data Mining Techniques

There are several techniques that are commonly used for data mining, such as:

- **Classification**

  It is most widely used techniques nowadays. It is based on the predetermined class to assign the sample into the related class. Applications of this technique are predicting of good or bad credit risk of loan applicant, and whether a patient has the certain disease. This is the technique that I will further on for this project. The details about classification technique will be made in the following review.

- **Association Analysis**

Also called market basket analysis. This technique used to generate models that produce the rules about combination of value-attribute, which will most likely to occur frequently together. For instance, rules that reveal the phenomena in which customers that purchased beer will more likely to purchase diapers as well. Thus, this analysis is beneficial for the marketers to arrange the combination of items together in the same market basket or having promotion for the items.

- **Clustering**

It is a descriptive technique that assigns a group to similar entities as well as to the dissimilar entities. In marketing, it can be used to find similar profiles for the patients and discover customer affinity groups. There are few methods used to implement clustering technique such as Kohonen net, demographic algorithms and k-means. Clustering is different from classification in the sense that we do not know about the label class of the group, and it is subjective. What is being employed in clustering is distance measure like which in the nearest neighbor technique. As a result, there will be different ways to cluster the data for data miners who are dealing with the same data sample. Therefore, clustering always need to have business domain expert to ensure the distance measurement is true.

Besides, there are methods that applied to the techniques in order to make the data mining a powerful approach to mine data from the massive databases. There are methods as below:

- **Artificial neural networks**: It is a non-linear predictive model that learn through resemble and training biological neural networks in structure.

- **Decision trees**: Sets of decision that represented by tree structure charts. This decision tree is then used to create classification rules for the dataset. Examples for decision trees are Square Automatic Interaction Detection (CHAID) and Classification and Regression Trees (CART) and Chi.

- **Genetic algorithms**: It is an evolution concepts based techniques that applied the processes such as mutation, natural selection and genetic combination that used in analyzing the biological approaches.
- **Nearest neighbor method**: Classifies the number of k data that are similar in the history data set into a class.

- **Rule induction**: The retrieval of if-then rules from dataset according to statistical significance.

Most techniques are widely used nowadays. It had been well applied in data warehousing as well as OLAP (On-Line Analytical Processing).

2.2.4 How Data Mining Work

Modeling is the main technique that used to make data mining works. What is modeling means? Actually, it is basically a model that created based on the known results and then applied the model to the data that we are going to analyze. With this model, others data can be examined without gone through any further of the past history dataset. Thus, it is quite 'handy' to analyze the data that are relevant to the model.

As the marketing analysis, we have lots of customers data about theirs age, credit history and the frequency of their shopping done. This is useful to find out the

potential customers and promote to them the items that they are more likely to purchase. With this, we know about the general information about the customers and prospects, and also the proprietary information for the customers. But, the proprietary information about the prospects becomes the target of our study. The main purpose of predicting is to make the General information to proprietary information for the customers. The model will facilitate in selecting targeted customers.

Besides, test marketing is a powerful source of data for this similar modeling. Finding the results form the market test, gives a foundation of identifying good prospects in the overall market.

2.2.5 Architecture for Data Mining

Data warehouse as well as flexible interactive business analysis tools integrated with data mining to make the full functionality of the data mining. Currently, many data mining tools are not operating within warehouse, thus extra process is required for importing, extracting, and analyzing the data. Moreover, with the integration with data warehouse, application results from data mining have been simplified when implementation is required. This can be practiced to improve the transaction process for the businesses, in the coverage areas of new product rollout, fraud detection, and promotional campaign management. Figure 1.1 illustrates architecture for an advanced analysis with large data warehouse.

Figure 2.1 - Integrated Data Mining Architecture

For an example of marketing-sales analysis tool, it consists of a data warehouse, which contains a combination of internal data tracking of all customer sales with external market data about competitor activity. Background information of potential customers also provided as basis for predicting. There is an advantage in which data warehouse can be implemented in a wide range of database management system. Examples of the DBMS are Oracle, Sybase, and Redbrick. It should support flexible and fast data access.

An OLAP (On-Line Analytical Processing) server is attached to enable a more sophisticated end-user business model. User can analysis the data based on the field they want to summarize. It may be summarizing their business by demography, regional, products and other key perspectives. The OLAP server must be integrated with the data warehouse and the Data Mining Server to drive in business analysis directly. Combination of Data Mining Server with the data warehouse enabies decisions implemented and tracked directly. When the warehouse deals with the new data, it can be used to mine the best practices and then apply them for future decisions making.

This shows a transformation of previous decision support system. It is not like the decision support system that merely sends the result to the user by query or reporting tools. With the Advanced Analysis System, business models had been applied directly to the warehouse before come out with a proactive analysis. The results of analysis provide a dynamic metadata layer therefore enhance the metadata of OLAP. Later, other tools such as reporting tools, visualization can be applied to make confirmation for the results.

Integration of customers, suppliers or marketers in a comprehensive data warehouse had caused exploration of information. Nowadays, there is an increasingly gap between the huge data warehouse and the ability of the users to retrieve relevant data for further analyzing. OLAP technologies and relational databases have capabilities to navigate immense data warehouse. Integration information system and the data mining will bring to a new era of information.

2.2.6 Data Mining Technique: Classification

Classification is a form of data analysis that used to extract models describing specific data classes in order to predict future trends. Generally, it used to predict categorical labels. Many classification methods can be found in expert system, machine learning, neurobiology and statistics.

In the classification model, case or record in the training data has been pre-classified as dependent variable. Training data from the training set are selected randomly from the data sample. As the label class is provided, it also called the

supervised learning. It is so call supervised learning as it is stated to which class sample belongs. It is contrasts with the clustering analysis in which the class label is unknown. In clustering, it is depends on the algorithm to group or class. There are some situations in which both techniques are used together. Firstly, we can use the clustering algorithm for grouping the similar samples together. Then, put the cluster into group and assign each sample to the cluster. Lastly, classification algorithm is used to find rules to assign new sample to the class in the future. Sometimes, the training data might be unknown, and it could be created by expert from this field.

It is important for a classification model to be able to apply in all possible outcomes in order to learn for other cases. For examples, in a loan prediction, cases for both good loaner and bad loaner should be included. In other marketing application, it is not applicable to the new prediction if the training data is available for customers of past promotions.

2.2.6.1 Dataset

Training dataset is dataset that used to build the model. Generally, the training dataset is gained from the databases, data mart or data warehouse. Test dataset is a sample data that withheld from the model-building process and used to test the model.

Datasets	
Training	Use modeling technique for model building
Control	Use to control for over-training (optional)
Test	Used by model builder for evaluating the accuracy of a particular model.
Validation	Used by data miner to evaluate the accuracy of final model by comparing predictions

Table 2.1: Functionality of datasets Function

Thus, training and testing datasets will always includes in the classification technique. However, it is possible to have more than two datasets that used to generate and validate a predictive model. For examples, the four datasets would be training, test, control and also validation datasets. The functions of each datasets are as below:

Is there a need to have so many datasets? As we know, data mining is an iterative process that has several levels of nested loops. At each nested level, a dataset is needed to properly test or validate the model. Data mining tool uses the training data for model building. Control data is used to select the best model from the various data models generated. After that, the model is tested. The training or control data cannot be used as the test dataset as it has been influenced those cases. Thus, the test dataset need to be form by independent data set. The learning parameters will be changed till the model builder is satisfied with the model obtained. However, in some cases the test dataset is overlap and be control dataset. Somehow, these datasets are recommended to be independent. For a dataset that is not used while model generation, it can be served as accuracy measurement.

Besides, n-fold validation can be used to determine the accuracy of a model. This is a method where all the data can be used to build the model instead of withholding partial of the data for model testing. But, the drawback of this method is many models need to be set up.

Model –building algorithm may use to evaluate a number of models. The collection of models then process through model-building algorithm to search the relevant model. The control set is used to select the best model by the algorithm whereas model builder uses the test set. Since the control and test set are used to select the best model, therefore, validation dataset is needed to validate the final output.

2.2.6.2 Sampling Techniques

The sample of each dataset should be independent variables which means that each of the sample only appear once in a dataset. There are few sampling techniques that used to avoid bias in the sampling of the dataset.

Random sampling is a widely used proper sampling. A sample is urged neither to take from the first and the last of the dataset, nor taken from the nth of data. With this, random number generator provides a method to select random sample. By using sampling, time consumed for training will be reduced. Different techniques of data mining required different passes of the model. For instance, the Naïve-Bayes technique needs only 1 pass through, whereas decision trees need several passes and for neural networks, up to hundreds or thousands of passes is needed.

Size of a dataset will have a great impact on development of data mining tools. Thus, for organization that have a massive data warehouse, they would prefer to build an algorithm that can run on parallel processor architectures. This can speed up through the training data.

For a very large dataset, it will cause high cost and administrative complexity. For a small dataset, likelihood of a seldom sample would cause the model to inaccurate due to the unusual data. Data mining is not distinguishable from noise. If the noise exists in the data, it will result in the following prediction. The accuracy of the predictions is not reliable. This type of incorrect predictions due to the existing of error in the training data is so call over-trained. Sampling can be used to solve many problems posed by large databases.

2.2.6.3 Accuracy: Model Measurement

To check the accuracy of the model-building process, testing and validation are used. It can be done either manually, or automatically. Usually the accuracy is tested before the model is used as predictor. This is done to validation dataset, which is dataset that does not share any data with training and control dataset.

Accuracy measurement of classification includes ratio of the correct prediction and the incorrect prediction. Confusion matrix is a more useful model that summarizes comparison between actual and predicted results. The results of the confusion matrices can be in the form of percentages or counts or both. However, accuracy is not sufficient to determine the usefulness of the model. Data mining like other decision-making or analysis tools is used to improve the business performance.

Therefore, many classification tools will include the summary that allows evaluation of the model's impact.

When predicting whether the customer is likely to respond to a direct mail offer, it is suggested to rank the prospects based on the probability, instead of just classify each candidate whether respond or not respond. This finding will later used to discover the highest ranked of the potential respondents. In order to analyze the possibility, a lift chart is shown as the graphical outcome. From the lift chart, any cut-off can be made to the desired response rate.

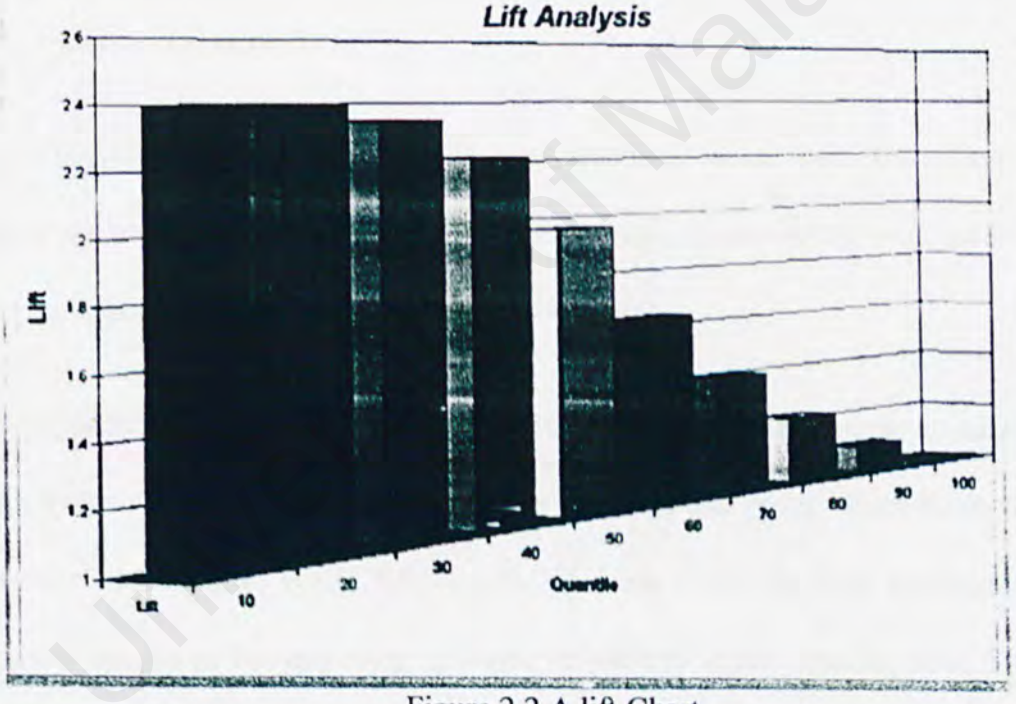

Figure 2.2 A lift Chart

Besides, lift information coupled with revenue and cost to determine the return on investment (ROI).

2.2.6.4 Predicting of New Sample

The common method of predicting is generate the output file from the input file with the output has an additional attribute, that is the predicting class or value of the sample. This can be applied through importing and exporting data to and out from DBMS.

Classification can be implemented through decision trees, Naïve-Bayes, neural networks and k-nearest neighbor. Classification technique should be applied more widely in the analysis tool. Besides, integration of data mining with databases should be enhanced to have a tighter coupling in order to retrieve data directly without repeated extraction each time analysis is needed.

2.2.6.5 Method: Decision Tree

A decision tree is a flow-chart-like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. (Han, 2001)

To classify an unknown instance, the attribute values of the sample are tested through the decision tree. It is traced from the root to a leaf node, which holds the prospects of the sample. While building the decision trees, the high numbers of branch will results in the appearing of noise or outliers in the training data. Tree pruning is used to overcome these problems by clearing the outliers of the branches.

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. (Han, 2001) Below is a sample of decision tree indicating whether a customer will buy a car:

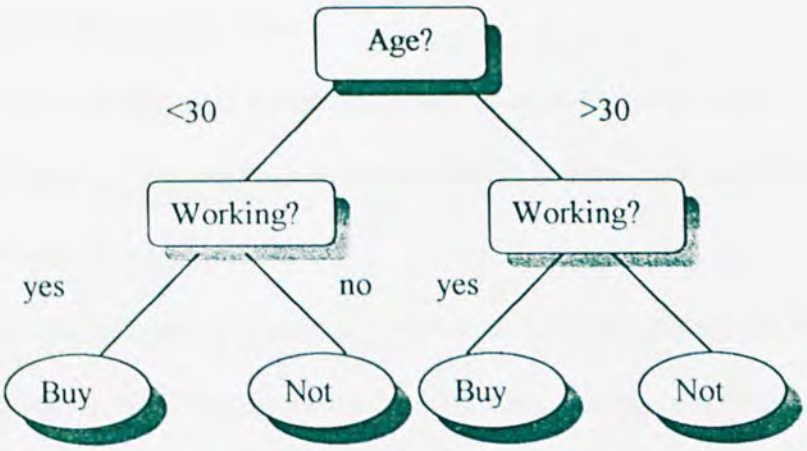

Figure2.3 A decision tree

The decision tree can be used to determine a group of customers whether they are going to buy car based on the previous training data.

2.3 Review of Similar Existing System

Nowadays, there are many available analytical tools in the market. Each tools had its function to generate information or knowledge from the raw data. In such a way, most manually efforts to extract relevant data can be minimized. This makes analysis an easier task compared to previous time. As data mining is significantly grow and yet it is not broadly used. Thus, while developing a data mining based analysis tools, we must consider about the user in which they might be not familiar with the new technology.

There are several similar existing systems that require the application of data mining.

2.3.1 STATISTICA Data Miner

STATISTICA Data Miner is a popular data mining software provider, which offers selection of data mining solutions on the market. It is icon-based and have extremely easy-to-use user- friendly interface.

With the completely integrated, automated and customizable features, It is ready to apply as business applications. The data mining solutions consists of 5 modules, such as: general Classifier, General Forecaster, General Neural Networks Explorer, General Modeler/Multivariate Explorer, General Neural Networks Explorer and also General Dicer Explorer.

Below are the features provided in STATISTICA software:

- **Data Warehouse**

It is defined as a process of organizing the storage of huge databases and facilitates to make the analytical process. STATISTICA software offers two-tier approach to store data. The first tier consists of simple structure of local files and it also can become component of second tier (provided in STATISTICA Enterprise System) It provides the data warehousing functionality. User can put up a variety of data types and unlimited size of datasets. STATISTICA is empowered to process extremely large datasets and even process the data from remote servers without importing to local database storage. Thus, processing datasets that are larger than local storage capacity also available. The second tier is offered in SEDAS (STATISTICA Enterprise Data Analysis System). It provides integration with other data repository and groupware functionality.

- **On-Line Analytic Processing (OLAP)**

It also called Fast Analysis of Shared Multidimensional Information (FASMI). It is a technology which users can generate descriptive summaries or

analytic results from multidimensional databases. It is not necessary to be in the real time environment, but it may use to analyze multidimensional databases, which contain dynamically updated information. It can be integrated to corporate databases and allow performance monitoring of the business. The results of OLAP would just be a frequency table, descriptive statistics, seasonal adjustment or removal of outliers.

- **Groupware**

Enable group of users on a network collaborates on specific projects. It may allow email communication, collaborative document development, reporting and tracking.

- **Scalable Software Systems**
- **Exploratory Data Analysis (EDA) and Data Mining Techniques**

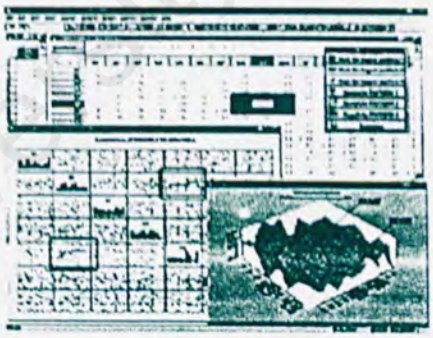

Figure 2.4 Sample of EDA, Hypothesis Testing and multivariate exploratory techniques

EDA is used to identify relations between variables, which there is no expectation to the nature of the relations. EDA uses a variety of techniques to search for the trends or patterns. For computational EDA techniques, multivariate exploratory techniques are used to identify patterns.

- Multivariate exploratory techniques identify datasets include: Cluster Analysis, Factor, Stepwise Linear, Nonlinear Regression, Function Analysis and Time Series Analysis, Classification Trees, General CHAID Models, General Classification and Regression Trees.

Figure 2.5 Neural Networks

- Neural Networks are analytic techniques available in STATISTICA modeled after process of learning system and neurological functions of the brain. It also predicts new observations from other observations after learning from previous data. Graphical data visualization methods allow identification of trends, patterns and relations.
- Brushing, an interactive method to select on-screen data points or data subsets therefore identify their characteristics and effects between variables. For instance, there is a technique of selecting all data points that belong to certain category from a matrix scatter plot. This is to examine the relations between variables and the observations. (See the subset in the illustration for the fourth component graph of the first row.) STATISTICA offers a comprehensive brushing technique, analytic brushing and interactive animated brushing by selecting attributes of specific data points.

Figure 2.6 Brushing

- Verification of the results is being done by serving the exploration of data as first stage and the not confirm results as tentative. For example, cross-validated which using a different dataset. It uses particular model from exploratory stage, and verify it by applying it to a new dataset and testing its predictive validity.

2.3.2 The Easy Reasoner (TER)

The Easy Reasoner (TER) software applied the data mining techniques such as decision tree, Case Based Learning (CBR), Rule-Based and K-Nearest Neighbor. It can be used in the Unix and Windows platform. Basically, it is a Case-Based Retrieval capable for the Eclipse inference engine product. Various machine-learning techniques such as inductive learning of decision trees (or rules) and nearest neighbor classification are being applied. CBR can use to classify new information and at the same time retrieve old information based on the distance between these information. After this, Eclipse rules can leverage the range of knowledge sources when adapting the retrieved cases to the new situation.

- **Basic Data Exploration Data Exploration through Visualization**

The Easy Reasoner classifies new information based on rules uses proven inductive techniques to construct decision trees. For a new record input, by traversing the decision tree, classification of the new data can be determined algorithmically. Retrieved records processed directly by end-users to complete the classification. Applying methods can be used to the similar cases. TER automatically learnt to classify. Besides, it also can retrieve using nearest neighbor techniques. In nearest neighbor retrieval, the distance between a new case and existing cases stored in a database is measured. TER also supports dBase databases. It can be explored with any third party tool that can view dBase database.

- **Support Model Building**

TER can build a decision tree with the Decision Trees and Nearest Neighbor Techniques for a set of databases. It uses API.

- **Data Mining Techniques Used: Supervised Induction**
- **Tool that handles missing data values/handle null available**
- **Weakness in this software:**

  It does not support GUI in this current version.

2.3.3 CART (Classification and Regression Tree)

Platform for the Software: Windows 95, Windows NT, Unix, IBM VMS & CMS, Windows 3.x

CART is an automation decision tree for classification and regression system based on data mining, data processing and predictive modeling. It automatically searches

for significant patterns and relationships from the databases. CART need let expertise to use the tool as it empowered with GUI, interactive Tree Navigator and intelligent default setting. It is reliable and stable performance proven. With the advanced features and batch production mode, it delivers versatility, accuracy and speed.

- **Discovery of Data Patterns**

It automatically discovers the relationships of cause-and-effect, significant patterns and predicts trends. The results are presented in a flow chart form as well as a tree-shaped diagram. It had visual aided output, which helps to easily see the hierarchical interaction of the variables. Therefore, simple if-then rules can be getting directly from the tree. It can investigate any classification related task, whether it is categorical variables or continuous variables.

- In the process of classifying, firstly, CART retrieves the data from the list of potential predictors. Then, focusing on the variables in the top ranking from the CART model. Thus, it can optimize the speed of the data mining techniques.
- **Few innovations from CART:**
    - Resolve the problem in how big will the tree be
    - Two ways binary splitting is being used.
    - Testing and tree validation automatically
    - Provides new method to handle missing values.
- **Model Building Support**

CART ensures no stopping rules can be depends on to reveal the optimal tree. Thus, one approach is introduced which is the notion of overgrowing trees and

then pruning back. In the searching for the patterns, the automatic validation procedures is essential to avoid over fitting, or mistaken of the patterns that only applied to the training data.

- Embedded with test disciplines to ensure the patterns will hold up when analyze the new data.
- Data Mining Techniques: Supervised Induction
- Data Cleaning Function available
- Data Transformations Executes Tool
- Provides Tools for handling Missing Data Value and Handling Null Values
- Model Evaluation Tools available
- **Future enhancement:**
    - Model Reporting Functionality, handling character variables and sampling weights.

- **Weaknesses found in the software:**
    - Without Basic Data Exploration through Visualization

2.3.4 S-PLUS Professional

S-PLUS Professional is advanced analysis software for data mining and for statistical modeling. It uses the combination of flexible data analysis environment with graphical user interface. Thus it is a user friendly and flexible analysis tool. Besides, it allows data imported from other sources like Excel, ASCII, SPSS and SAS. There are toolbars, dialogs and convenient menus. For Statistics, it includes generalized linear model, tree models, linear and nonlinear regression, linear and nonlinear regression and time series. Besides, customization is allowed through S

programming language in order to fulfill the analysis needs. Over 80 types of 2D and 3D graph can be selected. User can create intelligent graph interactively and also control every details in the graph and produce high quality results. Trellis graphics are used to reveal hidden knowledge in complex data.

Besides, there are feature that allows exportation of graphs to PowerPoint and Word for printing or presentation. It uses S language, which is designed for data exploration and visualization, programming with data and statistical modeling. S provides object-oriented environment to discover interactive data. S provides flexibility and extensibility. S System technology is the platform for S-PLUS product line.

- **Supports Visualization for Basic Data Exploration**

S-PLUS offers over 80 types of 2D and 3D plots types, conditioning plots and brush and spin techniques. It includes histograms, box plots, basic scatter and line plots, classification and regression trees and linear and nonlinear regression plots. Trellis graphics is introduced. It is a unique visualization technique that allows the discovery of how two variables change with variations.

- **Discovery of Data Patterns**

Provides cluster analysis, classification and regression trees, factor analysis, linear and nonlinear regression plots and Trellis graphics.

- **Model Building Support**

S-PLUS offers more than 3,000 data analysis function such as modern and robust techniques. There are generally 3 models includes predictive models, descriptive model and neural net extension.

- Predictive Model includes:
  - Classification: Classification trees, logistic regression
  - Regression: Linear regression, nonlinear regression, robust regression, ANOVA
  - Time Series Forecasting: ARIMA, AR and MA models, robust and classical filters and smoothers.
- Descriptive Models includes:
  - Association discovery
  - Clustering: k-means, model-based clustering, partitioning arould medoids, fuzzy analysis
  - Sequential patterns discovery
- Neural net extensions

- Data Mining Techniques: Association Discovery, Clustering Analysis Supervised Induction, Sequence Discovery and Visualization
- Data Cleaning Up Function

Provides complete set of data manipulation functions to merge, sub setting and transformation. Detection of outliers is available through graphical and computational summaries.

- Provides Data Transformations Tool
- Provides Null Handling/ Data Values Missing Tool

If missing values are found, it treated as special data type propagated through computations.

- Model Evaluation Performs Tool available.

2.3.5 Megaputer Intelligence's Data Mining Tools

Megaputer Intelligence's Data Mining Tools are designed to all structured data such as time series, Boolean data and categorical. There are three modules such as: PolyAnalyst 4.5, PolyAnalyst Knowledge Server and PolyAnalyst COM Objects.

In the software, it provides data mining algorithms such as: Cluster, Classify, Decision Tree, Find Laws, Decision Forest, Find Dependencies, Discriminate, Stepwise Linear Regression, Market Basket Analysis, Text Analysis, Transactional Basket Analysis, PolyNet Predictor, Summary Statistics, Memory Based Reasoning and Link Analysis.

The Classify algorithm allows assigning cases into different classes. It provides automation classification rule after predicting which classes the case belongs to. This is done by PolyAnalyst exploration engine. The accuracy for the testing cases is equivalent to the accuracy of the training data. The Classify engine determines the statistical significance of the generalized rule. Classify engine can use Find Laws, PolyNet Predictor or MLR as driving mechanisms.

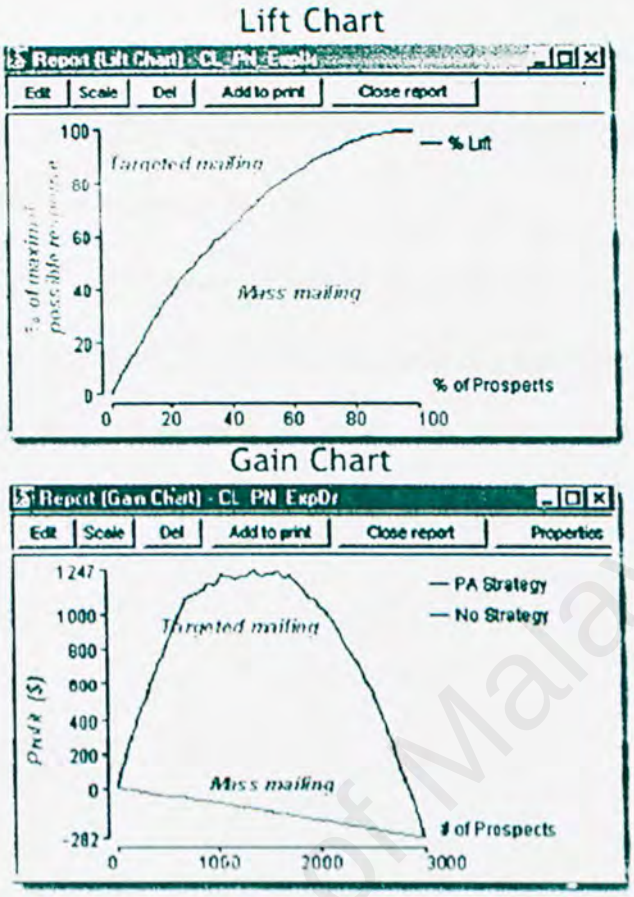

Figure 2.7 Examples for classify

- The PolyNet Predictor algorithm is a new combination to knowledge discovery in databases. It is a hybrid of Neural Net and GMDH (Group Method Data Handling). The new approaches of PolyNet Predictor can run on Windows NT and 95. It featuring a very effective and robust performance even dealing with large set of data.
- PolyNet Predictor generates an accurate prediction for the date through the examples from the previous learning. With functionality of classical Neural Networks, user can escape from building a whole proper network, due to the creation of a hierarchical structure of nodes dynamically by the software. Moreover, only one pass through of data is required for training purposes; therefore, training time of the network can be reduced.

- PolyNet Predictor offers two advantages include ability to construct arbitrary nonlinear models and it is efficient and fast to process large collection of data. Besides, PolyNet Predictor is suitable to use in cases when the target is to quickly predict the values of the data.
- The complexity of a hierarchical network structure and other characteristic built by PolyNet Predictor, are selected dynamically based on the properties of the analyzing data. If a built network is very simple, a rule expressed in Symbolic Rule Language could be obtained, which is equivalent to this network.

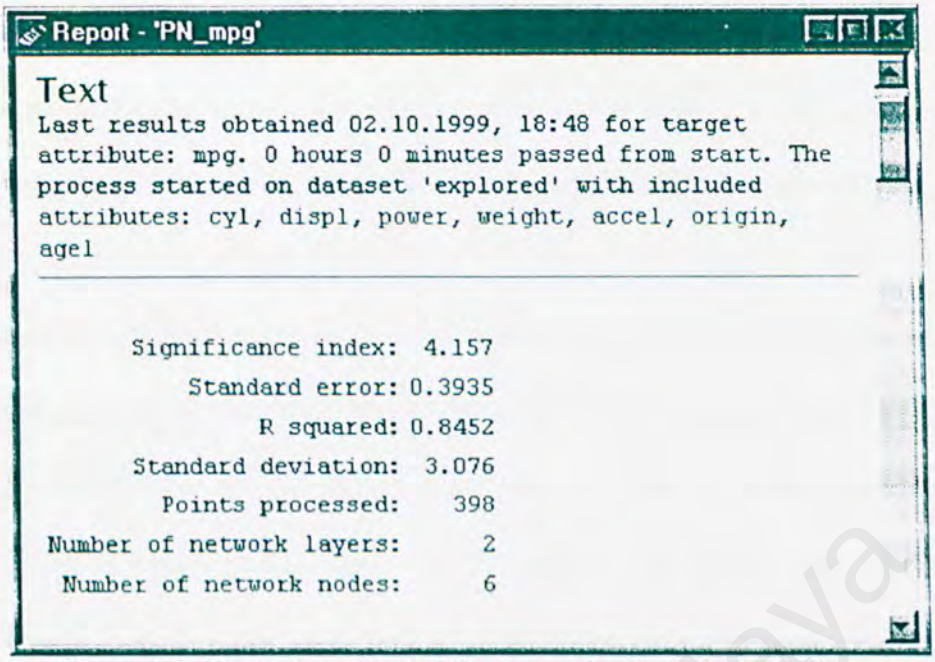

Figure 2.8 Report generated for neural network analysis

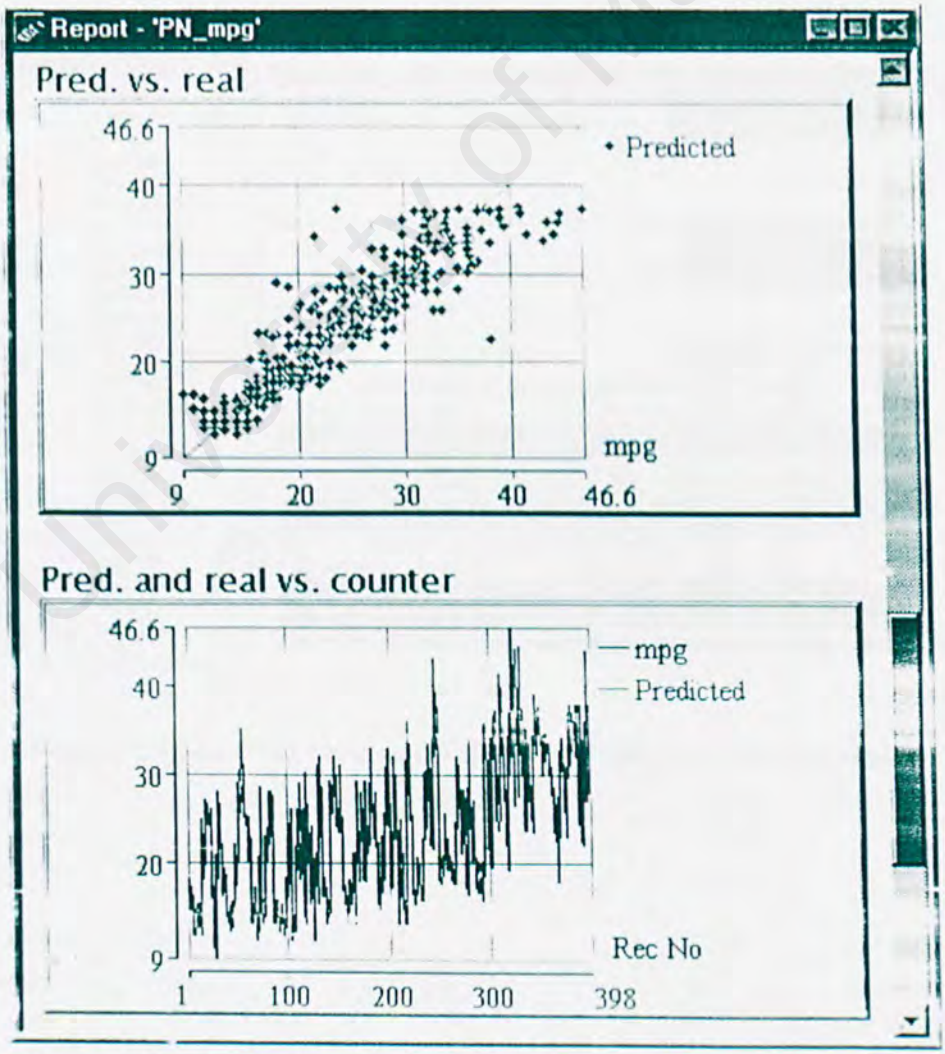

Figure 2.9 Results from PolyAnalyst Predictor

- Market Basket Analysis is an algorithm uses to examine a large list of transactions. The purpose is to determine which items are most likely or frequently purchased together. The results are beneficial to any businesses that involved in selling products in a store or a catalog, or direct sales. Basically, the main analysis in Market Basket Analysis is to find out, which products probably sell together. Thus, a list of sales transactions is used as input, where the data of the products and the sales or products are being analysis. If a product that is only sold few times of the entire, it should not be included. The algorithm will find rules involving these products that are not statistically significant.

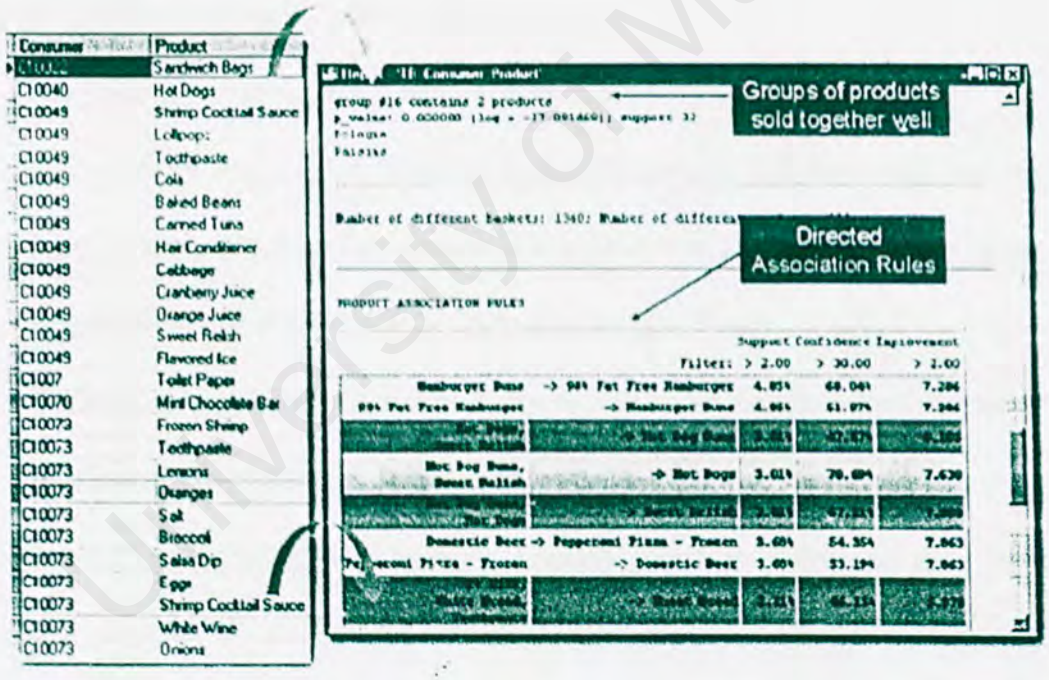

Figure 2.10 Market Basket Analysis that uses Association Rules

[DACS, 2001]

2.4 Review of Operating System

2.4.1 Windows NT Server

Windows NT Server can be considered as a remarkable network operating system. It does begin to meet remarkable commercial success. Windows NT is a cross-platform product, identical version are available for Intel X86, Silicon Graphics MIPS, and also Digital Alpha.

Windows NT Server has several features for networking. The most popular is the Distributed Common Object Model (DCOM). It is the combination features of ease-of-use of Windows 95 and the reliability of Windows. There are some advantages on this operating system such as below:

- User friendly interfaces that make the users ease to use
- Compatibility and productivity due to the Windows NT features that ensures high performance for 32-programs. It can be run in a separate address space responsively and reliably due to its multitasking features.
- Remote management and troubleshooting tools that enable administrator to implement policies and standards
- It protects critical device drivers, operating system code, and data from application. It runs the critical line-of-business programs. Separate allocations of the 16-bit applications as a fence to protect against bring down of other applications when failure occurs.
- Workgroup and networking supported. Built-in file sharing and print sharing make workgroup computing available. Besides, compatible open network system interface with other OS such as: Banyan Vines, NetWare, Novell,

Macintosh and Windows 95. For network environment, it support maximum of 10 simultaneous connections.

- Object Linking and Embedding (OLE) supported. Information from several Windows-based applications can be combined into one compound document.
- Built-in tools for internetworking and intranetworking. For instance, TCP/IP, Microsoft Internet Explorer and Microsoft Peer Web Services.
- COM and DCOM enable integrating of applications on a single computer or even across multiple computers.

2.4.2 Windows 2000

A lot of the momentum of computer networking is driven by the rapid growth of the Internet. Below are the features of Windows 2000:

- Includes Internet Information Services (IIS) 5.0 which features many improvements specially toward Internet Service Provider (ISP) who offer web hosting services.
- Supports HTTP compression, process accounting, quotas, and virtual server processor.
- Component Object Model (COM) allows applications to be updated centrally and distributed on the network. COM is now woven into just about every aspect of Microsoft's own software—operating systems, development tools, and applications. COM benefits both administrators and developers.
- Provides Active Directory, which it allows to locate any available network resource. It stores the location of objects and critical information about the object. For instance, a user's name, phone number, and address are all stored

within Active Directory and can be retrieved by anyone with proper access to the network. Many of Windows 2000 Server's network services store information within Active Directory to take advantage of its distributed, reliable nature. A resource as critical as Active Directory must be running at all times.

- Easier and faster administering a network of Windows 2000 systems. It is due to the Microsoft Management Console (MMC). Control over which computer and tools are displayed is given by the MMC. Thus, allowing you to create custom administration tools.
- SysPrep utility helps to reduce time it takes to build completely configured Windows 2000-based servers as compared to installing and configuring these servers by hand.

2.4.3 UNIX

UNIX is a popular operating system. Unix is used for workstations and minicomputers in the academic community initially, but now UNIX is available on personal computers. Business community has started to choose UNIX. PC and mainframe users are now choosing UNIX as their operating system.

UNIX, is a inter layer between hardware and applications.. It has both functions of managing the hardware and executing applications. Basically the differences of UNIX with other OS are internal implementation and the interface that seen and used by users. The UNIX users only need to be familiar with the interfaces and need not understand the overall internal workings. UNIX is an

operating system that includes the conventional operating system components. Moreover, a standard UNIX system includes a set of libraries and a set of applications.

There are two components in the hardware such as: the file system and process control with the set of libraries. On top are the applications. Normally, users have access to the libraries and applications. These two components make up the UNIX interface that may let the users think of as UNIX.

2.4.4 Linux

Linux is a freeware; which designed for Intel processors on PC architecture machines. It was written to avoid license fees, although the operation of the Linux operating system is based on UNIX. Linux shares the same command set as UNIX's, so one knows either UNIX or Linux, one will know the other as well.

Besides, there are few features of Linux such as supporting multitasking, multiprocessor, multi-platform, multi-user and multi-reading. Linux has protection for memory among processes; to make sure one program won't affect the whole system down. Those parts that are really used are only reads from the hard disk. It supports multiple processes to use the same memory to run.

Linux can link to either shared libraries (DLL's) or static libraries dynamically. User can debug on a program even when the program has crashed. Basically, it is compatible with System V, BSD and POSIX at the source level. Linux is unified memory storage for disk cache and programs in which all the

memory can be allocate for caching, and it reduces cache when a huge program is running.

It has POSIX job control. It supports various file systems such as Xenix, minix as well as V file systems. It offers a file system of up to 4 TB and the file name up to 255 characters. UMSDOS, which is a file system that allows Linux to be installed in DOS file system.

2.5 Review of Database Management System

Basically, there are 3 Database Management System (DBMS) taken to considering in this review. Those are Microsoft SQL Server 2000, Microsoft Access and Oracle.

2.5.1 Microsoft SQL Server 2000

Microsoft SQL Server 2000 is a product that meets the data storage and analysis requirements of the largest data processing systems and commercial Web sites. The same products can provide easy-to-use data storage and analysis services to an individual or small business. SQL Server 2000 features:

- A modern relational database engine that can scale from running on an individual desktop to running the largest Web sites. SQL Server 2000 is integrated with Microsoft Windows® 2000 fail over clusters to provide exceptionally reliable data servers, and integrated with Windows 2000 authentication and encryption to implement secure systems.

- Integrate with Microsoft data access environment. SQL Server 2000 provides native support for ADO, OLE DB, and ODBC. SQL Server also introduces integrated support for Web-based application development, supporting HTTP access using URLs, and returning data as XML documents.

- An integrated set of Analysis Services tools for performing complex data analysis and data mining of data warehouses.

- Replication services, which allow sites to place copies of data on multiple computers to improve overall system performance while keeping the data synchronized.

- Data Transformation Services (DTS) that make it easier to build OLAP data warehouses. DTS provides powerful services that allow records of individual transactions to be transformed into summary information stored in a data warehouse.

- English Query, which applications can use to answer ad-hoc user questions. When given a string containing a question about the data in a database or data warehouse, English Query returns an SQL or MDX statement that can be run to get the answer.

- Full-Text Search, which extends the pattern matching capabilities of SQL Server 2000 beyond the simple pattern matching available in the SQL language, including searches in files stored outside of SQL Server databases.

Meta Data Services, which provide facilities for storing, viewing, and retrieving descriptions of the objects in your applications and system.

2.5.2 Oracle

Oracle Corporation's reputation as a database company is firmly established in its full-featured, high-performance RDBMS server. With the database as the cornerstone of its product line, Oracle has evolved into more than just a database company, complementing its RDBMS server with a rich offering of well-integrated products that are designed specifically for distributed processing and client/server applications. As Oracle's database server has evolved to support large-scale enterprise systems for transaction processing and decision support, so too have its other products, to the extent that Oracle can provide a complete solution for client/server application development and deployment. Oracle Corporation has been a leader in introducing advanced client/server database technologies, directing its product development specifically to support the design, implementation, and management of client/server database systems. Oracle has designed products to support each of the three primary components of client/server architecture:

- A full-featured, high-performance RDBMS server, scaleable from laptops to mainframes
- Client development and run-time products that support multiple GUI environments
- Database connectivity middleware that provides efficient and secure communications over a wide variety of network protocols

Oracle's product offerings in each area are highly scaleable, providing complete client/server solutions for application environments ranging from small workgroup to global enterprise-wide environments.

2.5.3 Microsoft Access

Some features of Microsoft Access:

- Batch Updates in Access Projects by Using Microsoft Server 2000

  Access 2002 projects can batch all data entry and send it to the server when the user navigates from a record, closes a form, or selects a command. Besides, we can create a button on the form that saves all records or undo all changes to records, programmatically.

- Updateable Off-line Data Access Pages

  Data access pages in your Access project offline, make changes, and automatically synchronize when a reconnection is made to the SQL server. Changes to the off-line pages are made to an Access project connected to a local Microsoft SQL Server 2000 Desktop Engine (formerly MSDE).

- Password Security in an Access Project

  Logon password can be changed in an Access project connected to a Microsoft SQL Server 6.5 or later version database directly from within your Access 2002 menu.

- The Linked Table Wizard

The Linked Table Wizard guides you through the process of linking your tables to a SQL Server database, and does this all from within your Access project. It can delay loading. With this, software components that are not required for all databases, do not load into memory until they are needed.

- Improved call-tree feature enable modules to not load into memory until the Visual Basic code is executed.
- The analyzer Wizard analyzes the database, suggests the best way to maximize its speed and performance. Then, it will make the immediate changes after user had approved it.
- New filter for input enables you to search item in the database by input the filter analysis.
- Fail On Error property determines if an update or deleted query that is running will terminate if an error occurs.
- It can convert Macros to Visual Basic modules to perform equivalent actions.

2.6 Application Programming Language

Two application-programming languages are review here. That is Visual Basic and Java.

2.61 Microsoft Visual Basic 6.0

Visual Basic is a powerful development tool that exploits the key features of Microsoft Windows. It is famous with its ease to use features of the graphical interface. Besides, application can be developed in a short time.

Advanced database applications can be developed in order to access SQL Server database or any third party databases. This can be done by using visual database tools such as ODBC, DAO, RDO or ADO. Time consuming process can be reduced.

Besides, Visual Basic supports Graphical User Interface (GUI) that allows developer to enhance the interface design. There are few controls that help to create an easy-to-make interface. Several command buttons are ready to be used for the interface design.

Another advantage from Visual Basic is user can test and debug application within the development environments. Thus, Visual Basic is an interpreted language system.

2.6.2 Java

Java is a multipurpose programming language that introduced by Sun Microsystems. It is a language that can be used to build applications for the Internet. It is platform-independent. This means that it can run on computer running UNIX, MVS or Windows.

Java also enables the creation of applets to be embedded in a web document. Applets make graphics animation possible and together with the GUI, play audio clips.

Besides, it is designed to solve immense problems in programming practice. It can be used to develop advanced software for modern electronics. It support programming for real-world business solutions.

2.7 Summary

This chapter generally consists of three major parts. The first part is the review of the studies about data mining as well as classification technique. In this part, the data mining approach, usage, benefits and techniques is being discovered.

The second part is the review on the similar existing system. There are five analysis software been discovered here. From the review, the strengths and the weaknesses of the existing system become the reference to produce a better system as well.

The last part is the development tools review. This includes the review of operating system, programming language and Database Management System (DBMS). From the studies on the features of the tools, it helps me to further on to the System Analysis.

CHAPTER 3

SYSTEM ANALYSIS

Chapter 3: Methodology

3.1 Introduction

Before developing the system, an appropriate method had been chosen in order to give a direction towards the system output. Methodology is way how a system is developed from the very beginning since the analysis of the system requirement till the maintenance of the system. Thus, proper method or project model should be chosen depends on the system that we are going to develop. As for software development, there are several methods or development models such as Waterfall Model, Model V, Waterfall Model with Prototyping, Prototyping Model and so on.

3.2 System Development Model

For this system, the model that I applied is Waterfall with Prototyping Model. This is due to several advantages and reasons that it's suitable for this system. The reasons are:

- Every single activity can be examined after each step, based on the prototypes from time to time.
- Have the guidance and requirements that written in the prototype.
- From analysis till maintenance, every step in the process is guided with the prototypes and the testing will be easier.
- Help to assess alternative design strategies and decide which is best for a particular project.
- Easy to understand, thus easy to explain to the customer or user
- Design stages with the reference of prototypes make the process easier.

There are 5 main stages of waterfall model with prototyping. Each stage had been done with several methods in order to complete the stage and further to the next stage.

1. System Analysis and Requirement

The first stage is system analysis and requirements. From the analysis, few related languages, software, and also existing system had been analysis on the appropriateness to develop the system. Information and references for this related information are gathered through several sources, mainly Internet and reference books. Besides, some information and requirements are obtained through the sample theses from the senior projects. After gathering the relevant data, data had been analyzed and put it into the requirement document. Therefore, the requirement would be the guideline to be followed up throughout this development.

2. System Design

The second stage is System Design. After get to know about what is required in the new system, we can start on doing the system design. The system design will be based on the functionality and requirements. For this stage, the overall architecture of the system is established.

3. Implementation of the System

The third stage is Implementation for the system that had been designed. The program had been coded to set up the system according to the requirements.

4. Testing and Integration

The following stage after implementation is testing and integration where test for the developed system on the usability, functionality and robustness of the system. This is to ensure that the system is well developed before delivery. Besides, it also aims to make sure it meets the user requirements. Enhancement is needed to improve the system.

5. Operating and Maintenance

The last stage is maintenance, which is a long-term activity that started since the system launched. It is important to ensure the system remain useful all the while. Besides, further enhancement can be taken if that is a growing need.

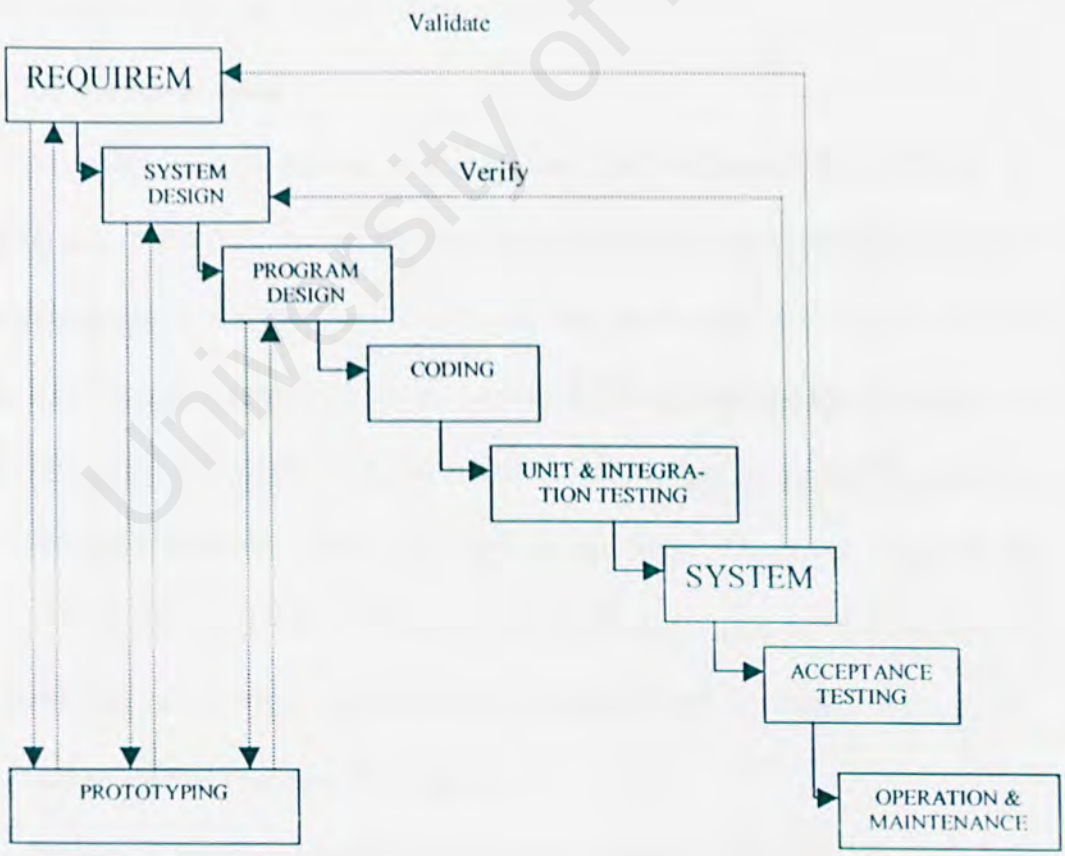

Figure 3.1 :Model Waterfall with Prototyping

3.3 Information Retrieval Methods

To prepare for this project, several ways had been tried to get the related and useful information of data mining as well as the development tools. This had been done towards primary data and also the secondary data.

- **Reference Book**

For the primary data collection, I had searched the related information about the data mining techniques and application from the reference book. Information from the book had given me idea about the new technology of data mining. Besides, the system development approach is also referred from the book as well. The information gained from the book is more theoretical and thus I also used other options to searched for the needed information.

- **Internet Searching**

Nowadays, search engine in the Internet had minimized the difficulty of searching task. From the Internet, the information about the latest release of software or techniques can be searched easily with only type in the related key words. This is the benefit of the information technologies. By searching in the Internet, the information gained is more likely to practical knowledge, in which the data are always depicted with the graphics as well as the demo. Moreover, some of the information is quite up-to-date. I found a lot or relevant data from the Internet not merely the text, but also some software that is useful for this development.

- **Interview with System Developer**

Besides, I took the opportunity to have an informal interview with system developer, Mr. Danny about some related software tools, development languages, and hardware. From the interview, I get to know about the latest popular language,

databases, software are being used nowadays. I also found what is suitable for my system development.

- **Analysis on the similar system**

A review on the similar existing system using the data mining techniques had been done in previous chapter. From this review, the strength and weakness are found in the existing system and I can enhance the existing system to create a better system. Some features had been added to make the system better and more user-acceptable.

- **Document Collection**

Documents that contain data and information about the functional area are gathered. The resources are gained from the library and the document room of FSKTM. I also get the report guideline from the document.

- **Faculty Website**

Some information are found in the faculty website that link to the thesis guideline.

- **Observational Research**

In observational research, I observed and statistically analyzed certain function and pattern that is being applied in the existing system in the market. Thereby, new principles or discoveries can be established. For example, some analysis tool can be used to determine the most desirable markets whereas some are used to predict the behavior of the loaner. Developing an objective system for quantifying observations is necessary to collect valid data.

3.4 Summary

A good method will lead to a better outcome. Thus, the methodologies that had been taken in this project are ways that relevant to get the information and sources I desire. I found that the methodology should be multi-directional in order to gain information from different sources, views and situation.

CHAPTER 4

SYSTEM ANALYSIS

Chapter 4: System Analysis

4.1 Introduction

Analysis of the system and user requirements is carried out in order to gather and interpret data, diagnose problems and use the relevant information for designing and developing the system. Through system analysis and development, we can identify problems, opportunities, objectives and analyzing the information system to solve problems. This had given us more knowledge for the respective field or system that should be developed in the future.

4.2 Target Group Definition

As data mining is an automation method to develop data filtering and interpreting of the system, thus, it's more relevant for the user from the decision making group. They will be the staff from marketing department, sales and so on. Besides, data mining is suitable for Database Administrator too, as they do not have to deal with the huge database to retrieve the data.

4.3 Analysis of requirements

A document of requirements captured all the features of the system or a description of functional and nonfunctional condition in which is able to fulfill the system's purposes. It describes the constraints on the system's performance and also the flow of information to and from the system. Requirement elicitation is the critical part of the process. Users' need and customers' requirements are determined

by several of techniques. Basically, requirement represents what would be the output of the system and the design identifies how to set up the system.

Requirement elicitation allows us to write a requirements definition documents of the system. Requirement definition is a complete listing of customer need of the proposed system. It represents an integration of needs and wants between developer and customer. Thus, the requirements definition of the sales-market analysis system is written based on the common requirements that are suitable for the system. On the other hand, the requirements specification restates the requirements definition in a more technical way for the system design development. It is the technical counterpart to the requirements definition documents and is written by the requirement analysts.

There are two categories of requirements such as functional and nonfunctional requirements.

4.3.1 Functional Requirements

For functional requirements, it contains what the users and customers ask for the functionality of the system. Functional requirements describe an interaction between the system and its environment. (Pfleeger, 2001) Each function had its role to fulfill user's or administrator's need. For the sales – marketing analysis tool, the functional requirements are divided to functions for User Module and the Administrator Module.

4.3.1.1 User Module

In the User Module, there are many functions that adapted to it. This is the module where user can access to the analysis system. Thus, the requirements of the user module are purposely set up for users to interact with the system. The users here refer to the sales officer, administrative clerk, management level officer and marketing personnel. The main functions that included in the module are:

a.) Input Function

Input function is to let the user input the appropriate data in order to access the information of the system. The purpose of this function is mainly to input the user's log in information and the field of data they are going to analysis.

b.) Display Function

This function will retrieve the results of the analysis and display in the suitable form of output.

c.) Save Function

After retrieving the analysis results, the user can have the option to save the results of the analysis. This is to make the user convenient to use the data in the future.

d.) Report Function

This function will show the result in tree decision model for the decision maker to further exploit the result.

4.3.1.2 Administrator Module

The administrator module is provided to allow the maintenance of the system. Administrator can easily access to it when modification is needed. Besides,

there is function that enforced the security of the system. The functionalities of this module are:

a.) Access Function

Administrator is given rights to make modification of the coding and the algorithm when it is needed. The sole of this function is to prevent unauthorized modification of the algorithm and coding.

b.) Maintenance Function

System developer or production support group member provides maintenance or troubleshooting when problems occur. This is also needed if the requirements changed.

4.3.2 Non-functional Requirements

Beside functional requirements, there are non-functional requirements where the requirements do not depend on the environment of the system. Somehow, this requirement play important roles as regulation that have to follow as to set up the system. There are several areas of coverage. A non-functional requirement is a description of the features, characteristics and attributes of the system as well as nay constraints that may limit the boundaries of the proposed solution (Pfleeger, 2001). It also defined as constraints under which the system must operate and the selection of the language, platform, tools or implementation techniques. The non-functional requirements are reliability, efficiency, accuracy, user friendliness, security, serviceability and usability.

4.3.2.1 Reliability

A system is considered as reliable if it should not cause unnecessary and unplanned downtime of the overall environment. Besides, it does not produce dangerous or costly failures when it is used in a reasonable way. With this, sales-marketing analysis tool must be ready when authorized user log in to the system.

4.3.2.2 Efficiency

Efficiency of the system is defined as the ability of a process or procedure to be called or accessed several times to produce consistent outcome in the acceptable speed. Thus, the system should have immediate reaction towards the user demand. In order to make sure the efficiency of the system, the algorithm of the technique must been tested to produce the accurate results.

4.3.2.3 Accuracy

Accuracy is the precision of computations and control. The accuracy is determined by the good data integrity, accurate databases and data consistency. Besides, it requires the predicted trend, and results are close to the real situation.

4.3.2.4 User Friendliness

The interface that does not require many "click" and not misleading user is considered as user-friendly interface. This quality is important as the interface is the inter communicator between the system and the user. A user-friendly interface is needed to reduce the learning curve for users. Therefore, the interface should be coherent as well as standardized.

4.3.2.5 Security

Security enforcement of the system is needed as the system only allows the authorized users to access the results and information. This is done through user log in to access the data and administrator log in to do the maintenance procedure.

4.3.2.6 Serviceability

This requires the system to be available all the times when users need to access the data. The requirement for this system is once the users log in, they can access to the data or results in less than 1 minute.

4.3.2.7 Usability

The system should be ease to use and accessible. Rather than limit or restrict the process, it should enhance and support the system. For this requirement, interface should be consistent and intuitive. Besides, the interface should be had less button, for instance, only 4 button each page, so that it won't mislead the user in selecting the function.

4.4 Development Environment and Tools

4.4.1 Operating System Platform – Microsoft Windows 2000

Windows 2000 had been selected as the platform for the system. It is based on several advantages and appropriateness of applied it to the system. Below are the advantages of Windows 2000:

- Includes Internet Information Services (IIS) 5.0, which features many improvements especially toward Internet Service Provider (ISP) who offer web-hosting services.
- Supports HTTP compression, process accounting, quotas, and virtual server processor.
- Component Object Model (COM) allows applications to be updated centrally and distributed on the network. COM is now woven into just about every aspect of Microsoft's own software—operating systems, development tools, and applications. COM benefits both administrators and developers.
- Provides Active Directory, which it allows to locate any available network resource. It stores the location of objects and critical information about the object. For instance, a user's name, phone number, and address are all stored within Active Directory and can be retrieved by anyone with proper access to the network. Many of Windows 2000 Server's network services store information within Active Directory to take advantage of its distributed, reliable nature. A resource as critical as Active Directory must be running at all times.
- Easier and faster administering a network of Windows 2000 systems. It is due to the Microsoft Management Console (MMC). Control over which computer and tools are displayed is given by the MMC. Thus, allowing you to create custom administration tools.
- SysPrep utility helps to reduce time it takes to build completely configured Windows 2000-based servers as compared to installing and configuring these servers by hand.

4.4.2 Database Management System – Microsoft Access

Due to the appropriateness of Microsoft Access to the proposed system requirements, it had been chosen as the database system to stored the data and control the data input or output. This is due to the features below that it's suitable for the sales- marketing analysis system.

- Password Security in an Access Project

  Logon password can be changed in an Access project connected to a Microsoft SQL Server 6.5 or later version database directly from within your Access 2002 menu.

- The Linked Table Wizard

  The Linked Table Wizard guides you through the process of linking your tables to a SQL Server database, and does this all from within your Access project. It can delay loading. With this, software components that are not required for all databases, do not load into memory until they are needed.

- Improved call-tree feature enable modules to not load into memory until the Visual Basic code is executed.
- The analyzer Wizard analyzes the database, suggests the best way to maximize its speed and performance. Then, it will make the immediate changes after user had approved it.
- New filter for input enables you to search item in the database by input the filter analysis.

- Fail On Error property determines if an update or deleted query that is running will terminate if an error occurs.
- It can convert Macros to Visual Basic modules to perform equivalent actions.

4.5 Programming Language – Java

As to develop the sales – marketing analysis system, Java had been chosen as the programming language. It is due to several features of Java that make the system more applicable. The advantages of Java are:

- A high-level programming language
- An object-oriented language that had been simplified to eliminate language features which cause common programming errors.
- Can run on most computers because Java Virtual Machines (VMs) (Java interpreters and runtime environments), which found in most operating systems, including UNIX, Windows, and the Macintosh OS. This is compatible for Windows Platform.
- Bytecode (Compiled Java source code file) can be easily converted to machine language instructions through just-in-time compiler (JIT).
- Suit well on the World Wide Web. It is compatible with common used web browsers such as: Microsoft Internet Explorer and Netscape Navigator. If the system is going to be set as Web-based system.

4.6 System Requirement

There are several requirements for the system. This includes hardware and software requirements.

4.6.1 Hardware Requirements

The hardware requirements are as below:

a.) A Processor with at least Pentium 166 MHz

b.) A minimum 64 MB RAM to support the workload

c.) A minimum of 1 G hard disk storage

d.) Other standard computer peripherals.

4.6.2 Software Requirements

The software requirements for the system are:

a.) Microsoft Windows 2000

b.) Microsoft Access 2000

4.7 JCreator

JCreator is a powerful IDE for Java. JCreator provides functionality such as : Project management, code-completion ,project templates, debugger interface, editor with syntax highlighting, wizards besides a fully customizable user interface. User can directly compile or run Java program without activating the main document first. JCreator will automatically find the file with the main method or the html file

holding the java applet, then it will start the appropriate tool. It is written entirely in C++, which makes it fast and efficient compared to the Java based editors/IDEs. Professionally designed to meet windows interface guidelines. The user interface will let users feel familiar

4.8 Tree Induction Algorithm

Tree Induction Algorithm

The algorithm operates over a set of training instances, *C*. If all instances in *C* are in class *P*, create a node *P* and end, else select attribute M and create decision node. Partition the training instances in *C* into subsets according to the values of information gain. Then, apply the algorithm recursively to each of the subsets *C* .

Choosing Attributes

Information theory is used in ID3 to determine the most informative attribute. The content of information is the reflection of probability of receiving the message:

Information (K) = 1 / probability (K)

It is then applied logs (base 2) to make information correspond to the number of bits required to encode a message:

Information (K) = - $\log_2$ (probability (K))

Information and Learning

The information content should be related to the degree of surprise in receiving the message Messages with a high probability are less informative than messages with

low probability. In the other way, learning is to predict a result accurately, which is to reduce surprise. To get the probability of two or more things happening, we take the multiplication of the probability. With taking logarithms of the probabilities, information can be added instead of multiplication. Level of uncertainty is:

$$- \Sigma \, p \log_2 p$$

To split the criterion, information gain for each attribute had been worked out to choose the best attribute.

Splitting Criterion

Assume there are *k* classes *Cm*, ... Cn

To decide which attribute to split on:

- Calculate the information gain that results from splitting on that attribute
- Split on the attribute that gives the greatest information gain.

Calculate the information gain from splitting *N* instances on attribute A:

- Calculate the entropy *T* of the current set of instances.
- **for** each value *a_j* of the attribute *A*(*j* = 1, ..., *r*)
  - Suppose that there are *J_j, m* instances in class *Cm*, ..., *J_j, n* instances in class *Cn*, for a total of *J_j* instances with *A* = *a_j*.
  - Let *q_j,m* = *J_j,m*/*J_j* , ..., q_j, n = *J_j,n*/*J_j*
  - The entropy *T_j* associated with *A* = *a_j* is -*q_j,m* log2(*q_j,m*) ... -*q_j,n* log2(*q_j,n*)

- Now compute T - $(J\_m/N)T\_m$ ... - $(J\_r/N)T\_r$ - this is the information gain associated with a split on attribute A.

Calculate the entropy T of the current set of instances :

- Suppose that of the N instances classified to this node, Im belong to class Cm, ..., In belong to class Cn
- Let $pm = Im/N$, ..., $pn = In/N$.
- Then the initial entropy E is $-p1 \log2(p1) - p2 \log2(p2) ... - pk \log2(pk)$.

[Ross Quinlan]

4.9 Summary

This chapter consists of the analysis of the proposed system in which the requirements needed to develop the system. Developing approach, database development tools, operating system, and programming language are determined in this chapter.

The requirements for the system are divided into functional requirements and non-functional requirements. Requirements are determined through the gathering of the information of existing system and the relevant to the system. It will be the guidelines to develop the system in order to reach the users' need.

To determine the appropriate development tools, it needs the pre-analysis of all the availabie option in order to get the most suitable tools that meet the requirements.

CHAPTER 5

SYSTEM DESIGN

Chapter 5: System Design

5.1 Introduction

A good system requires a good design. Thereby, the design of a system plays an important role in system development. From the system analysis and methodology that had been done previously, a proper system with suitable tools and languages can be developed. This includes designation of the system structure, database design, data flow diagram, and user interface design.

System design is important, as the design will results as the outcome in the end of the development. It means that, every steps of development have to match the system requirements whether is functional or nonfunctional requirements.

5.2 System Functionality Design

System functionality design is based on the requirements. The design translates the system requirements into system functionality. It represents in Data Flow Diagram.

5.2.1 Data Flow Diagram

The data flow diagram is drawn according to the Structure System Analysis and Design Method approach.

DATA FLOW DIAGRAM – LEVEL '0' DIAGRAM

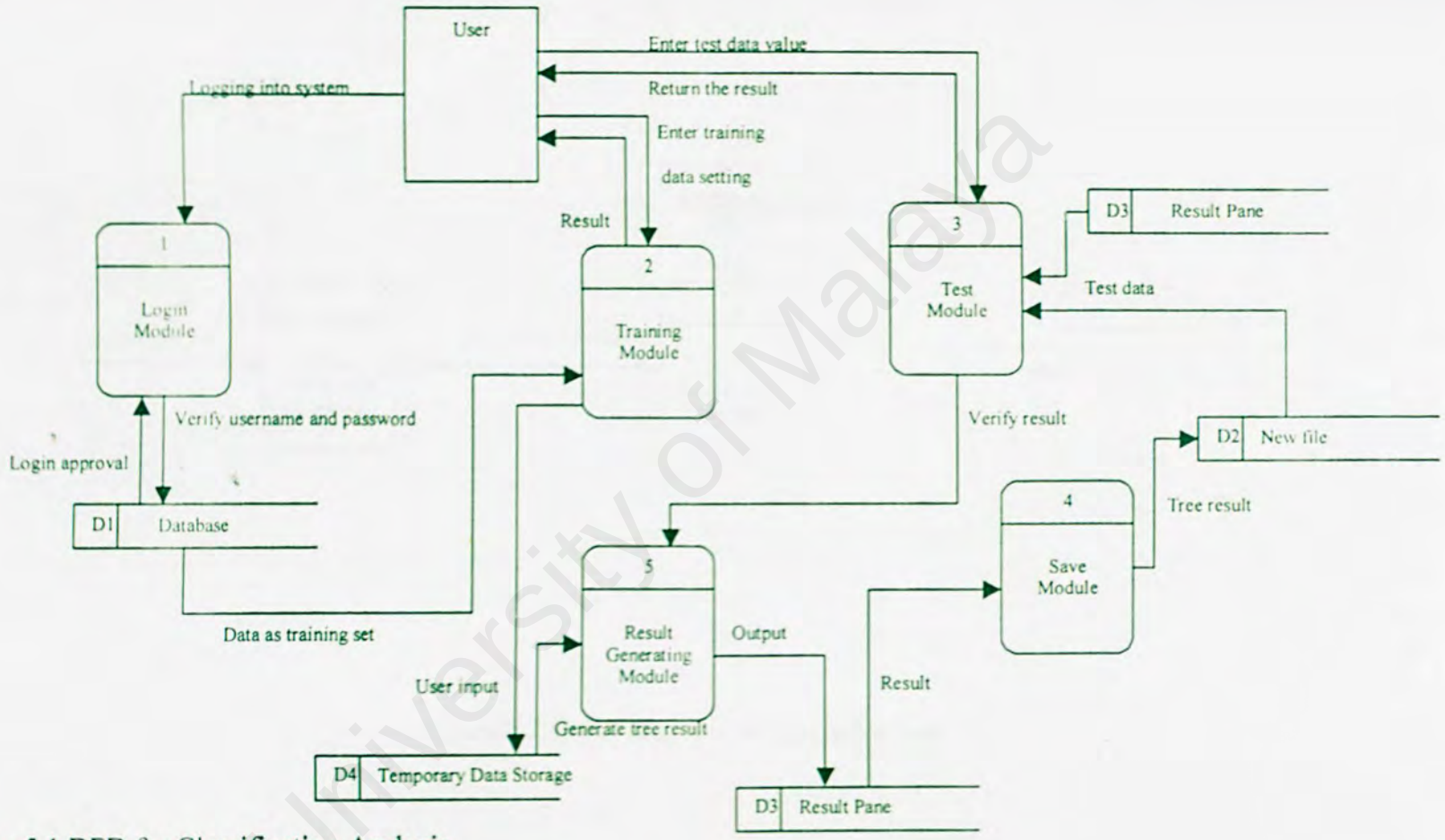

Figure 5.1 DFD for Classification Analysis

DATA FLOW DIAGRAM – CHILD DIAGRAM FOR PROCESS 1 (Login Module)

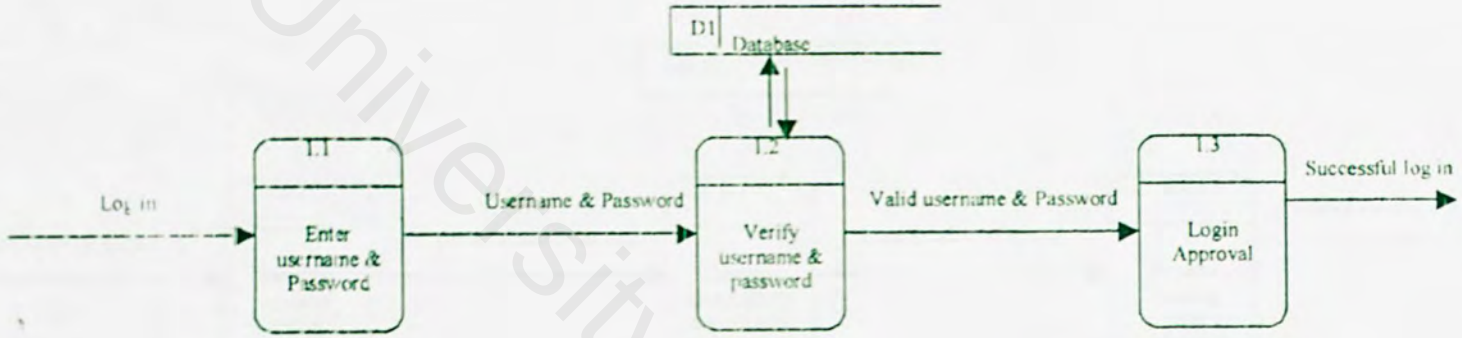

Figure 5.2 Child Diagram for Login Module

DATA FLOW DIAGRAM – CHILD DIAGRAM FOR PROCESS 2 (Classification Training Data Module)

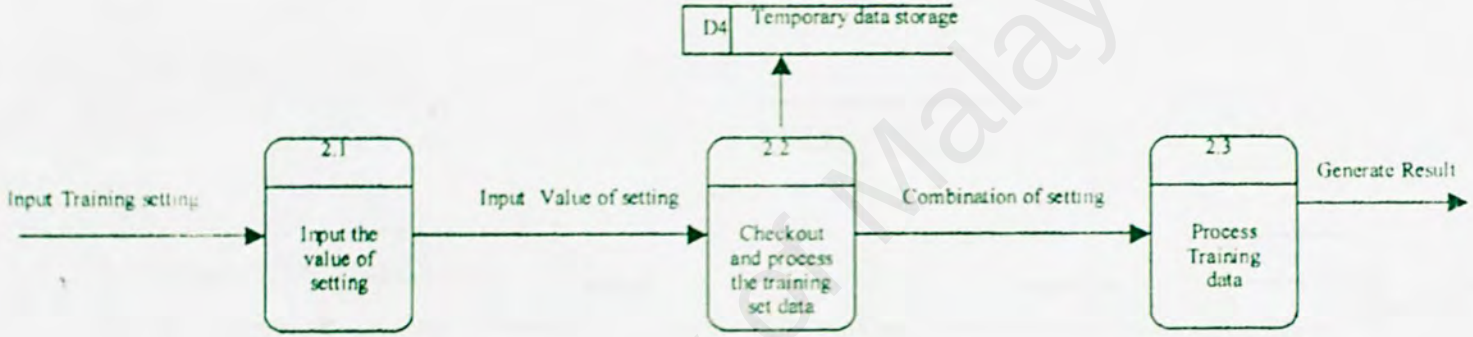

Figure 5.3 Child Diagram for Classification Training Data Module

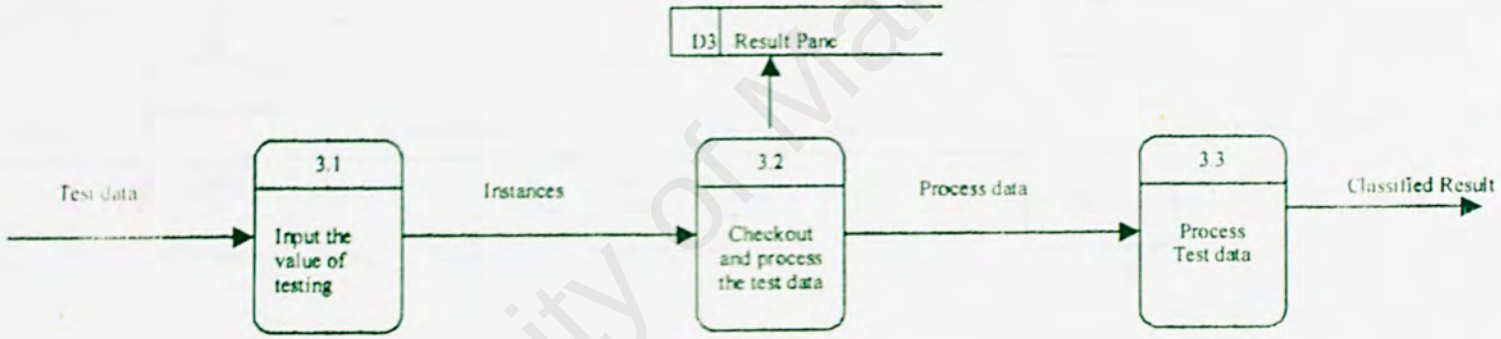

Figure 5.4 Child Diagram for Classification Test Data Module

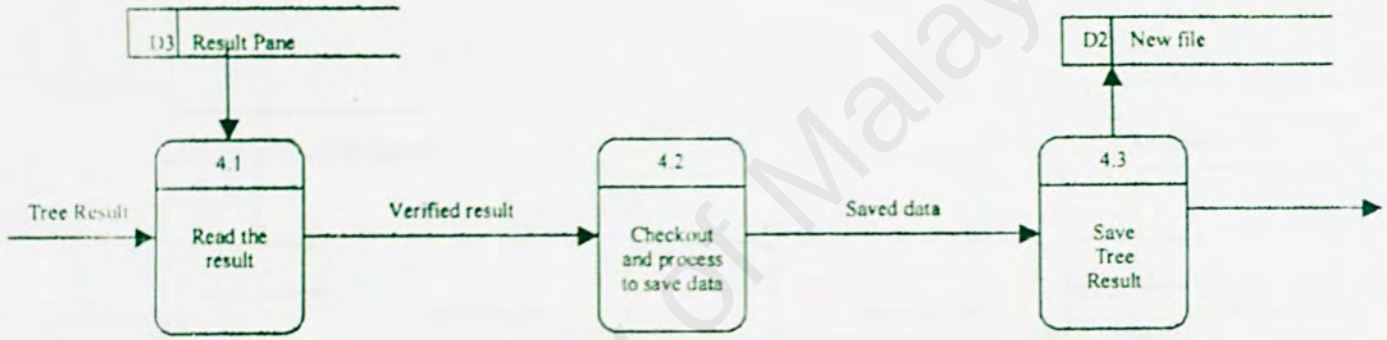

Figure 5.5 Child Diagram for Save Module

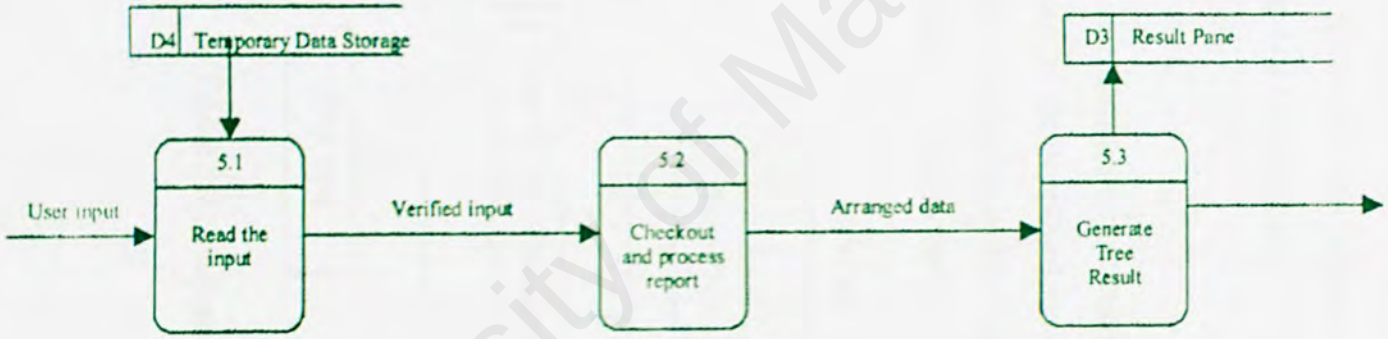

Figure 5.6 Child Diagram for Result Generating Module

5.2.2 System Structuring

The system functionality design is based on the requirements. The design translates the system requirements into system functionality. The functional design phase emphasizes on the data flow diagram.

Structured Chart

A structured chart is also known as Hierarchical Chart. It is drawn to show levels of increasing detail of the developed system. It is a breakdown of the main system's functions into smaller modules, thus enhancing the manageability, understandability, and integrity of the system.

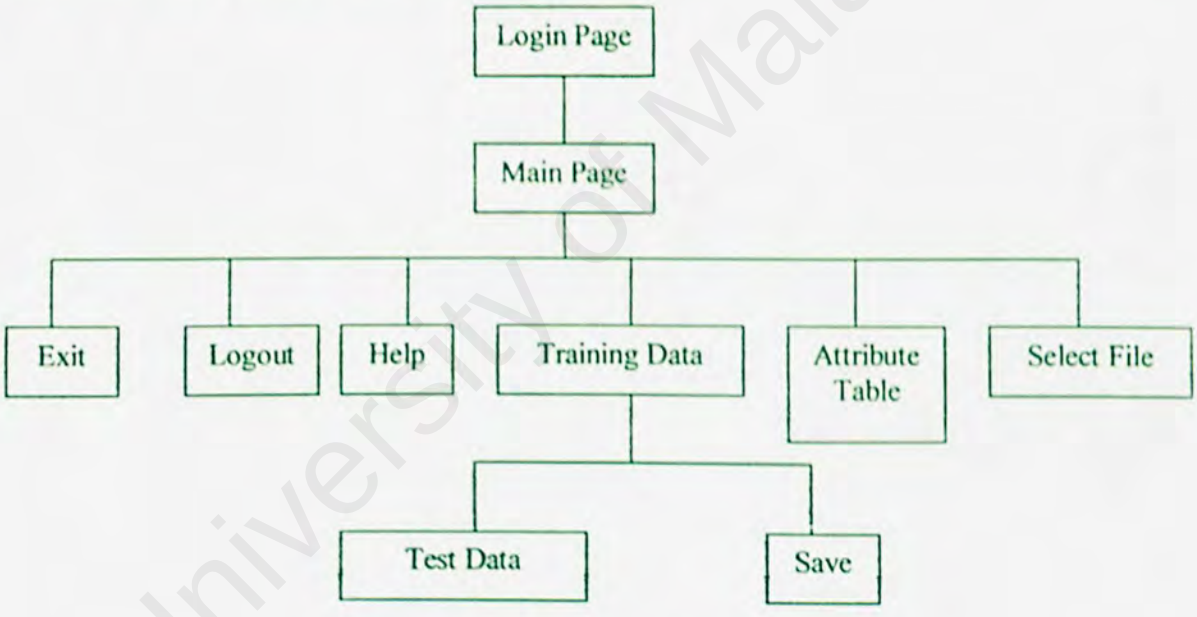

Figure 5.7 Structured Chart

LOGIN:

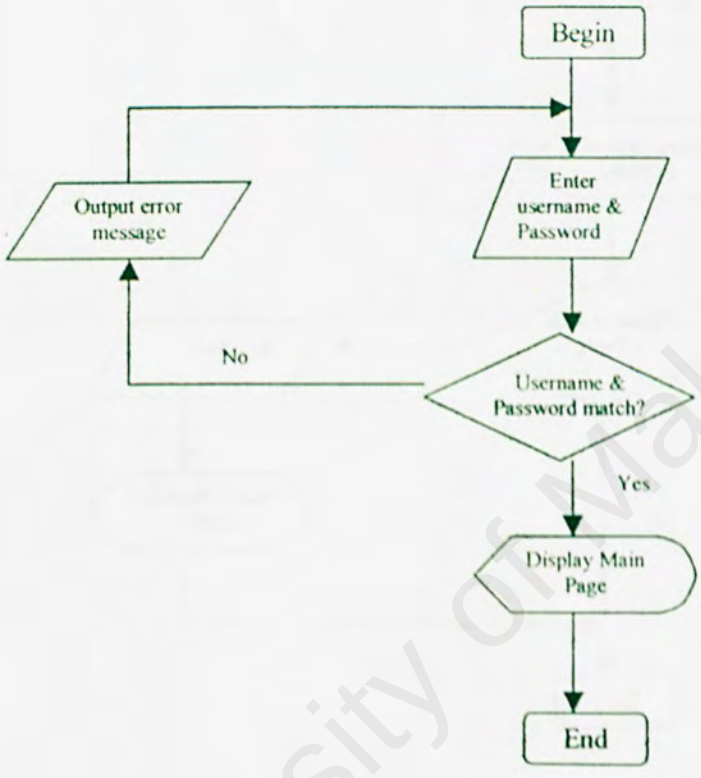

Figure 5.8 Diagram for Login

LOGOUT:

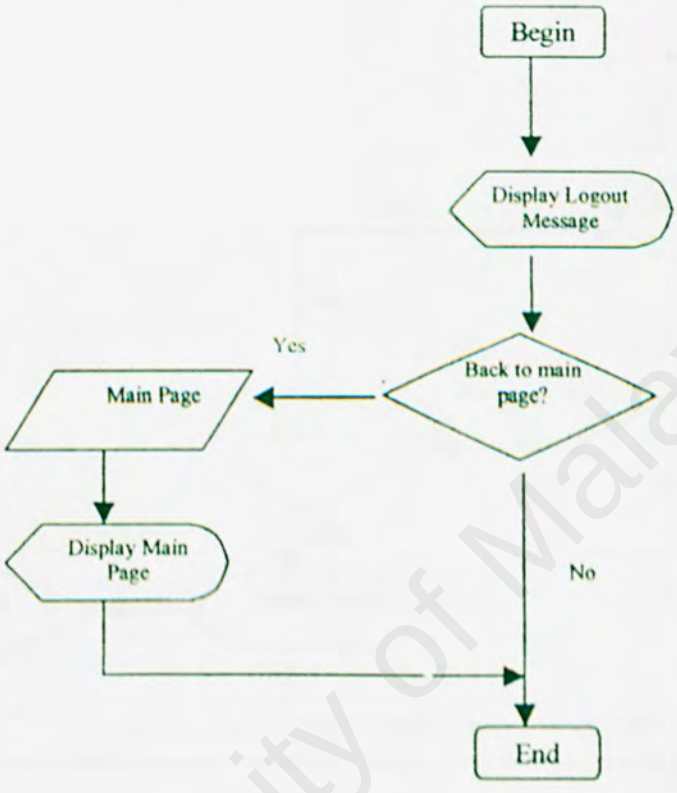

Figure 5.9 Diagram for Logout

TRAINING DATA:

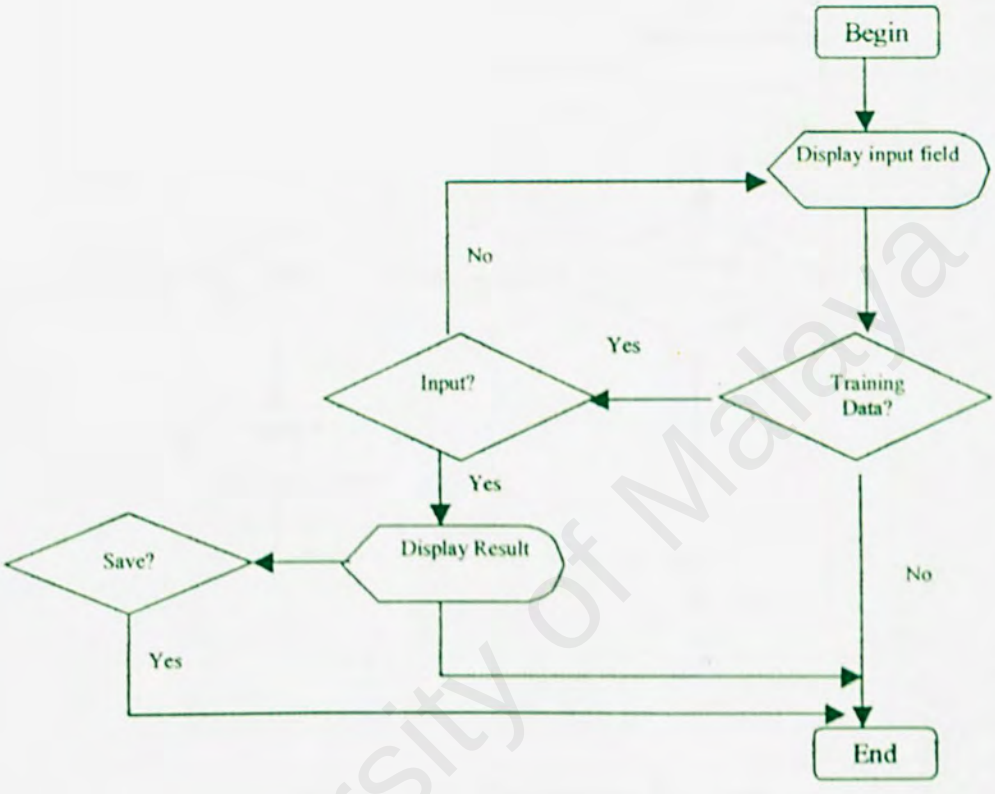

Figure 5.10 Diagram for Training Data

TEST DATA:

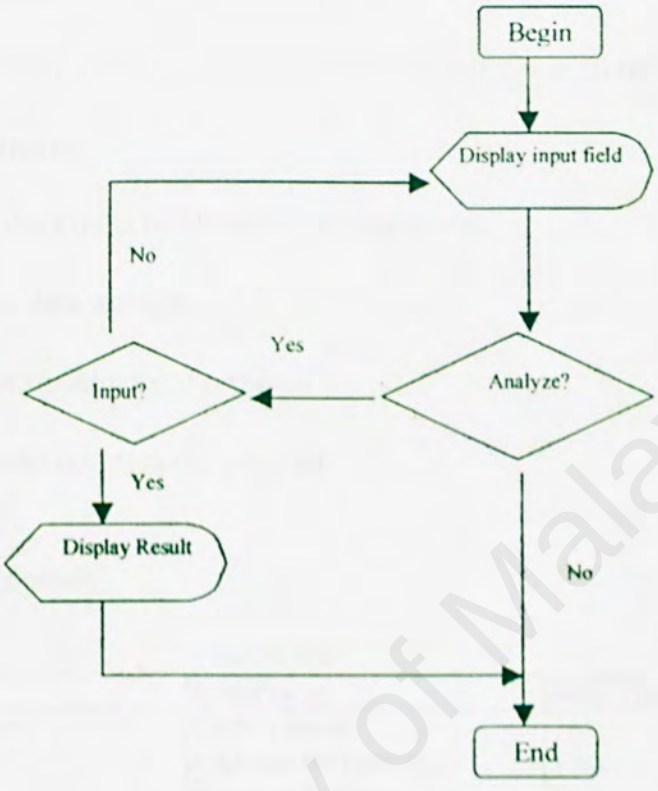

Figure 5.11 Diagram for Test Data

5.4 Database Design

The general objectives in the design of data storage organization are:

- Data availability
    - Data must be available when the user wants to retrieve
- Data integrity
    - Data must be accurate and consistent
- Efficient data storage
- Efficient updating and retrieval
- Purposeful information retrieval

5.4.1 Data Dictionary

Table:	PEMINJAM	
Field	Description	Data Types
ID_Peminjam	ID for Loaner	Text
Alamat	Address for Loaner	Text
Nama	Name of Loaner	Text
Kod_Pekerjaan	Code for Job	Text
Gaji	Salary gain for loaner	Number
ID_Cawangan	ID for bank branch	Text

Table 5.1 Table for Peminiam

Table:	KOD_KERJA	
Field	Description	Data Types
Kod_Pekerjaan	Code for Job	Text
Jenis_Pekerjaan	Types of job	Text

Table 5.2 Table for Kod Kerja

Table:	PINJAMAN_PAJAKAN	
Field	Description	Data Types
No_Pajakan	Index for Mortgage	Text
No_Pinjaman	Index for loan	Text
Status	Approval Status for loan	Text

Table 5.3 Table for Pinjaman_Pajakan

Table: PAJAKAN

Field	Description	Data Types
No_Pajakan	Index for Mortgage	Number
Jenis_Pajakan	Mortgage type	Text
Nilai	Value of the mortgage	Number
ID_Penjamin	ID for	Text
Nama_Penjamin	Name of	Text

Table 5.4 Table for Pajakan

Table: PINJAMAN

Field	Description	Data Types
ID_Peminjam	ID for loaner	Text
No_Pinjaman	Index for loan	Number
Jenis_Pinjaman	Types of loan	Text
Tarikh_Memohon	Date of applied	Date
Amaun_Pinjaman	Loan amount	Number

Table 5.5 Table for Pinjaman

Table: PEMBAYARAN

Field	Description	Data Types
No_Pinjaman	Index for loan	Number
No_Bayaran	Index for payment	Number
Peratus_Faedah	Interest rate	Number
Tarikh_Bayar	Date of payment	Date
Tempoh	Period	Number
Bayaran_Terkumpul	Accumulated Payment	Number

Table 5.6 Table for Pembayaran

Table: CAWANGAN

Field	Description	Data Types
ID_Cawangan	ID for branch	Text
Nama_Cawangan	Name of the branch	Text
Telefon	Telephone number	Number
Lokasi	Location of the branch	Text

Table 5.7 Table for Cawangan

Table:	CEK	
Field	Description	Data Types
No_Bayaran	Index for payment	Number
No_Cek	Index for Cheque	Number
Amaun	Amount of payment	Number

Table 5.8 Table for Cek

Table:	TUNAI	
Field	Description	Data Types
No_Bayaran	Index for payment	Number
No_Tunai	Index for Cash	Number
Amaun	Amount of payment	Number

Table 5.9 Table for Tunai

5.5 User Interface Design

The quality of system input determines the quality of system output. It is a vital that input fields and screens be designed with this critical relationship. Well-designed input fields should meet the objectives for:

- Effectiveness
- Consistency
- Simplicity
- Accuracy
- Ease of use
- Attractiveness

The user interface design is based on GUI. Besides, Human-Computer Interface general principles of designing an attractive system have been applied, such as:

- Consistency – Consistent format for command input, data, display, menu selection and placing of the control objects

- Confirmation and verification message – Ask for verification of any non-trivial destructive action such as delete records
- Recoverability – Ability of the users to take corrective action once an error has been recognizes
- Reverse Action – Allow user to return to previous state
- Functions Grouping – Categorize activities by function and organize screen geography accordingly
- Responsiveness – The rate of communication with the system

Interface Design:

Login Page:

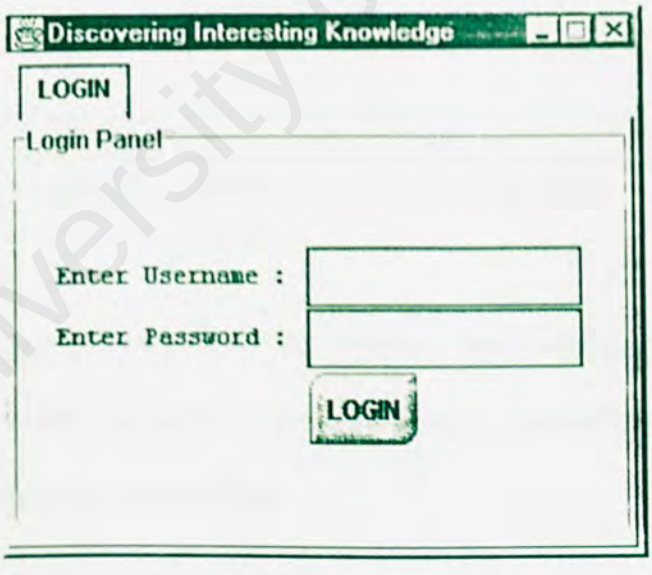

Figure 5.12 Interface Design for Login Page

Main Interface:

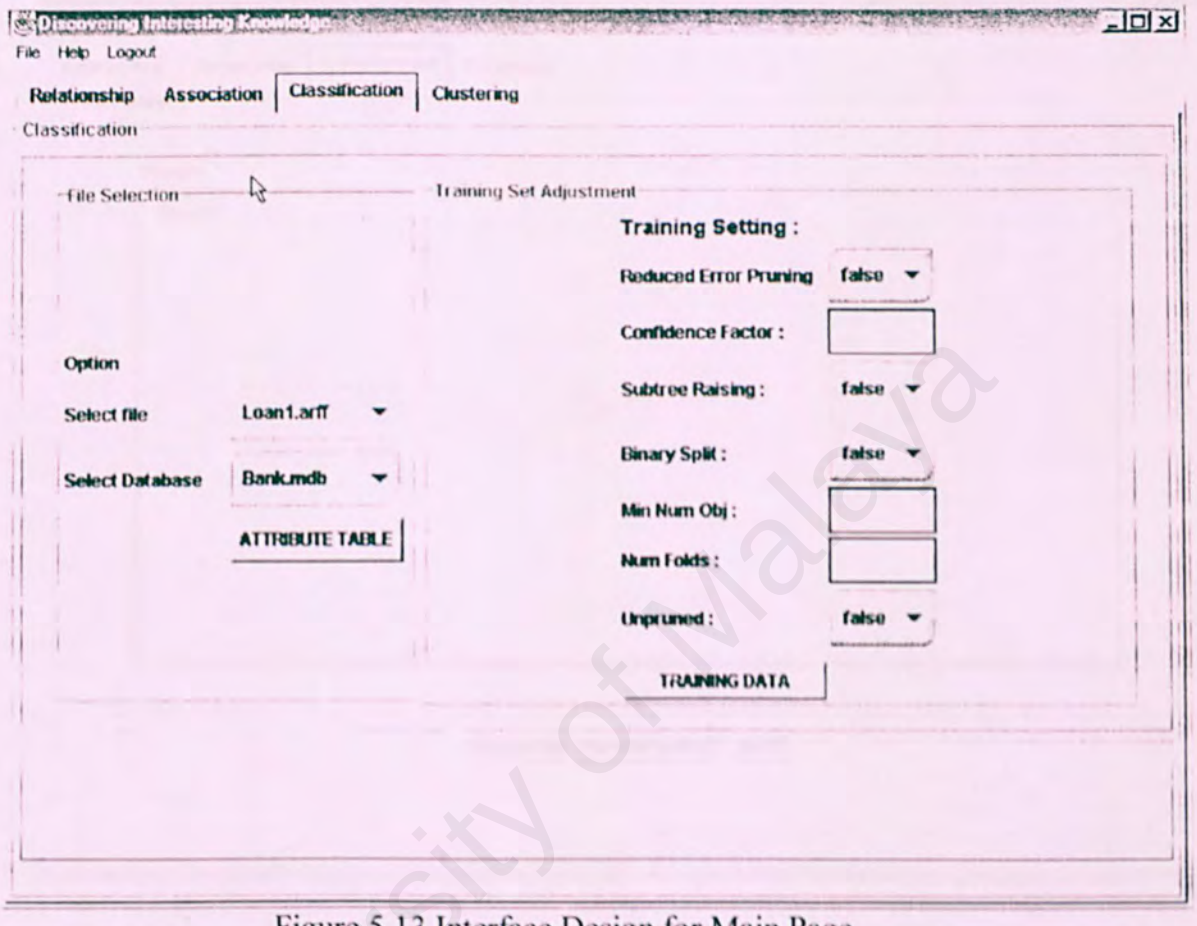

Figure 5.13 Interface Design for Main Page

The interface have two parts which are file selection and training set adjustment. User select the fiel to anaylze and then can set the training requirement. For example, using pruned or unpruned tree.

Result Page:

Figure 5.14: Interface Design for Result Page

5.6 Summary

System design is an important step to design the requirements of the system. This includes database design, system functionality design and user interface design. With this design, the development proceeds to next phase, system implementation.

CHAPTER 6

SYSTEM IMPLEMENTATION

Chapter 6: System Implementation

6.1 Introduction

System implementation is the stage of development after system design. It is the stage of transformation where plans in design transform to real system. In this data mining sales marketing analysis tool, implementation system is divided into module, which is platform development and module implementation.

6.2 Platform Development

To begin the implementation, we must have the entire requirement platform set up. For this system, there are operation system, databases and java editor that need to be set up before implementation of the system can be done.

6.2.1 Operating System Implementation

The first stage of implementation is to set up Windows 2000. This is done by reformatting the hard disc and makes it available for new operating system. Besides, to enhance the security system, NT File System had been set up. This is continues by installing Windows 2000 using installer. After the installation, the hardware configuration has been done.

The hardware used to develop the system which is stated in system analysis is as below:

- 450 Mhz Pentium Processor
- 128 MB SDRAM
- 10.0GB Hard Dick
- CD ROM

- 1.44 Floppy Disk
- Other standard desktop PC components

6.2.2 Database Implementation

Microsoft Access is used as Database Management System. It is used to store data for training data and user name and password. Thus, the software had been installed after the installation of Windows 2000. After installing Microsoft Access, I started to design for the database. It is a sales-marketing database in a bank environment. Since java is the programming language, ODBC-JDBC Bridge is used to link to database from the application.

6.2.2 Java Development Kit

Java Development Kid 1.3.1 (Jdk 1.3.1) is used to develop java program. With this kit, we can use predefined function in Java.

6.2.3 Development Tool

To write the program for java, we can either compile and run it in command prompt or we can used other option by writing it in the editor such as JCreator, JBuilder, JForte and so on. This editor facilitates the writing of coding and can easily trace the error and search for the error location. In this project, development tool used is JCreator. This is because JCreator is flexible and can used in any platform. It is easy to use.

6.3 Module Implementation

The development is implemented module by module. For each module, a prototype is created and tested. This prototype only had some function of the module. After that, other additional function added into it until functional and precise prototype is created. Finally, the entire modules are combined and integrated to form the application. Thus, module is created one by one; for example, log in module, save module and also the main classification module.

6.4 Standards and Procedure To Write A Code

Standards and procedures will facilitate to organize our thoughts and prevent mistakes. Thus, documentation is important to keep track of our system. There are some procedures to be followed whether it is internal documentation or external documentation. This is not only facilitating the process of writing coding, but also make the maintenance task easier. A standardized documentation will help in determine the error and locate them in order to correct it. While transforming design into code, standards and procedures are useful. With this, changes in design are easier due to the correspondence between design components and code components.

6.5 Program Documentation

Program Documentation consists of internal documentation and external documentation.

6.5.1 Internal Documentation

Internal documentation is documentation embedded in a program. For instance, there are internal comments in the program, which provide a clear and

understandable reference for the third party on what the particular coding. To the developer, comments communicating with other readers of the source code. Thus, I have enclosed comments in every function and in places where make reader confuse. Statements of purpose showing the function of the module and a descriptive comment, which embedded in the body of the source code, describe processing function.

Header Comment Block is usually located at the beginning of each component. In the header comment block, there are few information such as (Pfleeger, 1998):

- Component name
- Writer name
- Location of component in general system design
- Time component was written and revised
- Way the component uses data structures, algorithms and control

Program Comments is additional comments to help the reader understand how the intensions are implemented in the code. Besides providing a line by line explanation of what the program is doing, the comments can also break the code into phases that represent major activities. (Pfleeger, 1998).

In java, // and /* */ can be used to comment the particular coding.

For example,

- // Save result in the result pane
- /* Get all the input from user

*/

With this comment, I can read refer to appropriate line easily and so do the reader.

A Meaningful and Understandable Variable Names and statement label is given in order to make the coding readable.

For example,

```
JComboBox jComboxNF;
JPanel jPanel;
```

Compare with :

```
JComboBox a;
JPanel b;
```

The previous is more readable.

Besides, formatting is used to enhance understanding. Indentation and spacing of statements reflect the basic control structure (Pfleeger, 1998). For example,

```
class ButtonHandler implements ActionListener
{
        public void actionPerformed(ActionEvent ev)
        {
                String s=ev.getActionCommand();
                if(s=="LOGIN")          //Login is Clicked
                {
                    accessDB();

                }
        }
}
```

With this indent and spacing, can prevent error or reduce it to the minimum level.

Documenting data. Program readers find it difficult to understand the way in which data are used and structured. A data map is useful in interpreting the code's action, especially when a system handles many files of varying tyoes and purposes, coupled with flags and passed parameters (Pfleeger, 1998).

6.5.2 External Documentation

External documentation is all other documentation. The purpose of this documentation is to let the user who may never look at the actual code. Besides, more explanation can be done clearly in this documentation. It is a full-blown report which answers the questions on how, when, why, where and who to use the system. (Pfleeger, 1998).

6.6 Program Algorithm

Before transform design into coding, algorithm is needed as a guide to ensure every step is being coded. This is important, as it make sure there is a focus point during programming. The algorithm for this program is listed as below:

<u>Login Page</u>:

1.0 START

2.0 Display Member Login Page

3.0 If user wish to login

- 3.1 Enter username
- 3.2 Enter password
- 3.3 Click Login button to submit information
  - 3.3.1 If Login successful
    - 3.3.1.1 Display Main Page which is Tab Pane showing relationship diagram
  - 3.3.2 If Login failed
    - 3.3.2.1 Display dialog message indicating the error that has occurred.
    - 3.3.2.2 If user wish to try again

3.3.2.2.1 Go back to Login Page

4.0 END

Main Page:

1.0 START

2.0 Display relationship diagram

3.0 Display option in Tab pane

3.1 If user select Association

3.1.1 Go to Association page

3.2 If user select Classification

3.2.1 Go to Classification page

3.3 If user select Clustering

3.3.1 Go to Clustering Page

4.0 END

Menu:

1.0 START

2.0 Display Menu Item

2.1 If user select Exit from File Menu

2.1.1 Prompt dialog message, ask confirmation of exit

2.2 If user select Help from Help Menu

2.2.1 If user select Classification

2.2.1.1 Display Classification help page

2.2.2 If user select Association

2.2.2.1 Display Association help page

2.2.3 If user select Clustering

2.2.3.1 Display Clustering help page

2.3 If user select Logout

2.3.1 Prompt dialog message, confirmation of logging out

3.0 END

HELP:

1.0 START

2.0 Display Classification Help page

3.0 If user close the help window

3.1 End

4.0 END

Classification Page:

1.0 START

2.0 Display File selection

2.1 If user select file n

2.1.1 Use file n

2.2 If user select database

2.2.1 Use selected database

2.3 If user click Attribute Button

2.3.1 Display Attribute Table

2.3.1.1 If user close Windows of Attribute table

2.3.1.1.1 End

3.0 Display Training Set Adjustment

3.1 If user click Training without any input

3.1.1 Prompt warning message

3.2 If user input non-numeric character

3.2.1 Prompt warning message

3.3 If user input number between 0-1 for confidence factor

3.3.1 Prompt warning message

3.4 If user input number less than 0 into num folds and min num obj

3.4.1 Prompt warning message

3.5 Else

3.5.1 Go to result pane

3.6 END

Result Pane:

1.0 START

2.0 If user select Training Button

2.1 Go to Training Set

3.0 If user select Save

3.1 Prompt save window

3.2 Save the result pane

4.0 If user select previous page

4.1 Go to previous Classification Main page

5.0 END

Training Set:

1.0 START

2.0 If user select Analyze without input value

2.1 Prompt warning message

3.0 If user select Analyze with non-numeric character

3.1 Prompt warning message

4.0 If user select Analyze with empty column

4.1 Prompt warning message

5.0 If user select Cancel

5.1 End

6.0 If user select Clear

6.1 Clear both the input column

7.0 END

Logout:

1.0 START

2.0 If user wish to logout

2.1 Click logout

2.1.1 Display user logout message

2.1.2 If user wish to login again

2.1.2.1 Go to login page

2.1.3 Else

2.1.3.1 End

3.0 END

6.7 Summary

Implementation is a transformation of design to the real system. It consists of the method use to implement the design. From hardware configuration, software selection and documentation, every step is stated in the text. Documentation is important to produce readable and referable program. Algorithm of the system helps to make sure every flow of system is on the track.

CHAPTER 7

SYSTEM TESTING

Chapter 7: System Testing

7.1 Introduction

System testing is the phase of development after implementation stage. The main purpose is to ensure the system work under the proper way in which it should be. It is a critical element of software quality assurance, which performed to represent the ultimate review of specification, design and coding for this data mining sales-marketing analysis tool. The system is developed using Waterfall model, thus several interactive testing is generated to test the module. Every module in the system is tested and enhanced repeatedly to ensure precise functionality in the future. After every module had been going through unit test, the modules then are integrated. The integrated system then tested in integration testing.

7.2 Unit Testing

Unit testing is done with the intention to find out the errors and faults that beneath the modules. In this Data Mining sales marketing analysis tool, the module testing had been done through out the testing stage. Few areas had been focused on such as the source code, user interface and situation running. Thus, in testing the modules, methods such as casing testing, user testing and source code examining.

7.2.1 Source Code Examining

Source code examining was used to go through the coding to search for the error factor, which will lead to the mis-functional of the system later. This step is

imperative, as it will like the snowball principle that caused the bad consequent for the output.

The examining process is done by comparing and determining which part in the code is differ from the previous design of the module process flow. Through this process, the correctness of the source code would be found. For future tracing, comments were inserted in the sections as to make a clear reference.

As the development for java coding editor is JCreator, the errors in the code can be easily traced as the line number where the error falls would be shown. Besides, the other editing tools such as "Find", "Replace" and so on facilitate on viewing the particular code and make modification as well.

7.2.2 Test Cases

In order to test the module practically, several test set were designed to test the modules. Those cases were various range of input that could be input. For example, the modules were tested with all possibility of input item. With this, potential error and fault had been found when the circumstances changed. This includes the sample of training set data from the database. In this case, extreme inputs were tried and the response on the system was listed down for further checking.

Besides, the testing set function in the module had been gone through times over times to make sure the algorithm and the output make sense. In order to make sure the error message prompt in the proper way, several wrong inputs were applied purposely. Through out this method, the error message would only appear when it needed. This won't cause confusion to the user in the future.

Immediate Window facilitates testing directly by just assigning different arguments.

7.2.3 User Testing

After test cases were applied to the modules, it was still not considered as precisely simulated module. Thus, involvements of other users are needed, as they will provide different aspect of view to the module. Apparently, the user interface whether is user friendly or the contrary could be found out easily.

Besides, users will rise out some conditions, which I didn't notice when I was doing the testing for my own. The users consist of team members, course mates, friends and others. Each module was shown to them and let them have the first hand try on the module for their own. At the same time, the effectiveness of these modules in improving user learning curve also tested.

7.3 Integration Testing

When all the modules had been passed over the unit test, it would be integrated to become a whole system. Sometimes, problem occurs when those modules integrated into one main functional system. With this, we can ensure the components of the system would support one another. The objectives are: to compare the whole system with the functional and nonfunctional requirements, to detect fault or bugs in the integrated system and to examine the correct flow of the integrated system.

The development of this data mining sales marketing analysis tool is starting from modules and form by the combination or integrated of all in the end. In this testing stage, bottom up approach was applied, as it is most suitable for integration testing method.

7.4 Summary

After several times of testing, the testing stage is completed. Thus, it is reliable and can be used by user.

CHAPTER 8

SYSTEM EVALUATION & CONCLUSION

Chapter 8: System Evaluation & Conclusion

8.1 Introduction

Upon the completion of this project, system strengths and limitation had been verified and evaluated. With this, the evaluation by comparing the outcome and the requirements in the earlier stage of project had been found. These include the system strengths, limitations, problem encountered, objectives achieved and few future enhancements had been notify as well.

8.2 System Strength

As this is an analysis tool using data mining, thus the analysis result would be quite accurate since it's based on the algorithm for each technique. It is a program that generates the result for sales-marketing related decision in a bank environment. As for classification, a simple loan approval classifier had been made. With this, the bank manager can base on this as the first level of approval to the loan application. There are few features of the system that is useful:

- User-friendly interface. User can easily used this system, as it contain of standard graphical user interface control objects such as buttons, text box, dialog message and other window-based features. These make the user interface more self-explanatory and attractive.
- User can select file to analyze it using classification method. Thus, several cases can be analyzed and this allows time-to-time update of the training set result.

- User can select using pruned or unpruned tree method to train the data. This feature allows user to analyze the data using different selection. This allows different option for user if they wish to include outliers in the study.
- Save function allows user to save the result and review anytime. As for reference and up-to-date checkup, this is useful as user can base on different result in a period to see the trend of the loan approval.
- Login feature enables better security of the system in which authorized user can log in to the system.
- User can know better about the database through relationship of the database of the main page.
- Besides, through the attribute table, all the attributes available for data training is showed in the table.
- User can select the training data adjustment before generate the training result such as: number of folds, confidence factor, min number objects and so on. This is to improve the result of training set.
- Test data features used to generate the result according to the training set result. By inputting new combination of attribute, system will classify it and conclude it in the belonging class.
- Dialog messages that help user to have the right way to use the system especially where user need to input the value.

8.3 System Limitation

Somehow, there are still some limitations in this system.

- The system is limited to analyze files that are in arff format.

- There are still some functions need to be added.

8.4 Future Enhancement

In order to improve the features of Sales-marketing Data Mining Analysis Tool, few enhancements could be added. Below are some enhancements proposed:

- Automatically generate the arff from the attribute selected by user. This is to let user analysis whatever they want.
- Select other databases.
- Have the visual- aided result such as graph.

8.5 Problem Encountered

Upon completing of this project, I have encountered some problems. It might be unfamiliar with new language, design of the system, applied of data-mining concept and algorithm in the system.

Thus, several ways had been done to find the solution for problem-encountered.

Like what I have discussed in the methodology part, search the information through Internet, and reference book and asking help from friend which has great experience in this field, help a lot.

8.6 Objective Achieved

As overall review, this system had met the requirements and objectives, which planned at the early stage. User can analyze the sales-marketing environment data, accurately.

8.7 Conclusion

As a conclusion, this project is meaningful to me. This is because; I have the whole participation on this project from the planning till set up the system. Thus, I have come along the development of the system. Through out this experience, I gain something that is out of the book. Undoubtedly, experience gained through industrial training also helps a lot.

Somehow, this system still has something need to be enhanced in the future. This system had inspired me to do better in the future.

References

Data Mining: What is Data Mining
(http://www.anderson.ucla.edu/faculty/jarson.frand/teacher/technologies/palace/datamining.htm) [3 July,2002]

Doug Alexander
(http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex/) [July, 2002]

Estelle Brand and Rob Gerritsen: Classification and Regression
(http://www.dbmsmag.com) [June, 2002]

Han, Jiawei, Micheline Kamber. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Knowledge Discovery
(http://www.knowledgediscovery.com/) [13 May, 2002]

Knowledge Discovery Database
(http://info.gte.com/~kdd/) [11 May, 2002]

Mary Schrader: Microsoft Access
(http://business.baylor.edu/Mary_Schrader/OFF_XP/Access_XP/New.htm) [August, 2002]

Megaputer Intellige's Software
(http://www.megaputer.com/) [12 July, 2002]

S-PLUS Professional
http://www.mathsoft.com/splus/ [June, 2002]

Statistical Software
(http://www.statsoftinc.com/sys_integration.html) [July, 2002]

Stat Soft Inc. – Data Mining Software
(http://www.statsoftinc.com/textbook/stdatmin.html) [June,2002]

Stat Soft Inc. – Data Mining Software
(http://www.statsoftinc.com/datamine.html) [June,2002]

The Haley Enterprise
(http://www.haley.com/) [July, 2002]

Pfleeger, Shari Lawrence. (1998). *Software Engineering Theory and Practise*. Prentice-Hall Inc.